国外电子与通信教材系列

PEARSON

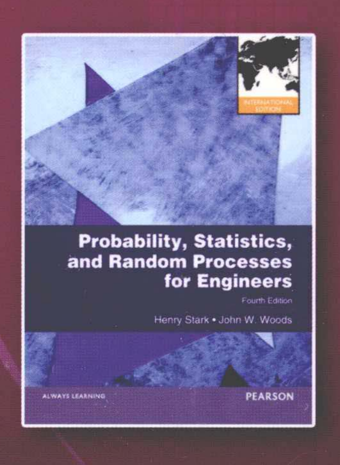
英文版

概率、统计与随机过程

（第四版）

Probability, Statistics, and Random Processes for Engineers

Fourth Edition



[美] Henry Stark
John W. Woods 著



http://www.phei.com.cn

概率、统计与随机过程（第四版）（英文版）

Probability, Statistics, and Random Processes for Engineers

Fourth Edition

本书从工程应用的角度，全面阐述概率、统计与随机过程的基本理论及其应用。全书共11章（其中第10章和第11章为网上资源），首先简单介绍概率论，然后各章分别讨论随机变量、随机变量的函数、均值与矩、随机矢量、统计（包括参数估计和假设检验）、随机序列、随机过程基础知识和深入探讨，最后讨论了统计信号处理中的相关应用。书中给出了大量电子和信息系统相关实例，每章给出了丰富的习题。

本书适合作为电子信息类专业本科生和研究生的“随机信号分析”或“随机过程及其应用”课程的双语教学教材，也可供从事相关技术领域研究的科技人员参考。

相关图书

本书中文版





www.pearson.com

For sale and distribution in the mainland of China exclusively(except Taiwan, Hong Kong SAR and Macau SAR).
此版本仅限在中国大陆发行。



策划编辑：马　岚
责任编辑：马　岚
责任美编：李　雯





国外电子与通信教材系列

概率、统计与随机过程

（第四版）（英文版）

Probability, Statistics, and Random Processes for Engineers

Fourth Edition

[美] Henry Stark
John W. Woods 著

電子工業出版社
Publishing House of Electronics Industry
北京・BEIJING

内容简介

本书从工程应用的角度，全面阐述概率、统计与随机过程的基本理论及其应用。全书共11章（其中第10章和第11章为网上资源），首先简单介绍概率论，然后各章分别讨论随机变量、随机变量的函数、均值与矩、随机矢量、统计（包括参数估计和假设检验）、随机序列、随机过程基础知识和深入探讨，最后讨论了统计信号处理中的相关应用。书中给出了大量电子和信息系统相关实例，每章给出了丰富的习题。

本书适合作为电子信息类专业本科生和研究生的“随机信号分析”或“随机过程及其应用”课程的双语教学教材，也可供从事相关技术领域研究的科技人员参考。

Original edition, entitled **Probability, Statistics, and Random Processes for Engineers, Fourth Edition**, 9780273752288 by Henry Stark, John W. Woods, published by Pearson Education, Inc., publishing as Pearson International, Copyright © 2012 Pearson Education Limited.
All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage retrieval system, without permission from Pearson Education, Inc.
China edition published by PEARSON EDUCATION ASIA LTD., and PUBLISHING HOUSE OF ELECTRONICS INDUSTRY, Copyright © 2012.
This edition is manufactured in the People's Republic of China, and is authorized for sale and distribution in the mainland of China exclusively (except Taiwan, Hong Kong SAR and Macau SAR).

本书英文版专有出版权由 Pearson Education（培生教育出版集团）授予电子工业出版社。未经出版者预先书面许可，不得以任何方式复制或抄袭本书的任何部分。
本书在中国大陆地区生产，仅限在中国大陆发行。
本书贴有 Pearson Education（培生教育出版集团）激光防伪标签，无标签者不得销售。

版权贸易合同登记号 图字：01-2012-5644

图书在版编目（CIP）数据

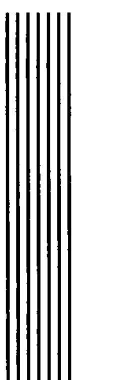
概率、统计与随机过程：第4版 = Probability, Statistics, and Random Processes for Engineers：英文 /（美）斯塔克（Stark, H.），（美）伍兹（Woods, J. W.）著. — 北京：电子工业出版社，2012.8
国外电子与通信教材系列
ISBN 978-7-121-17668-5

Ⅰ.①概… Ⅱ.①斯… ②伍… Ⅲ.①概率论 – 高等学校 – 教材 – 英文 ②数理统计 – 高等学校 – 教材 – 英文 ③随机过程 – 高等学校 – 教材 – 英文 Ⅳ.① O21

中国版本图书馆 CIP 数据核字（2012）第161886号

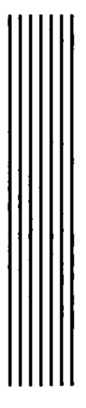
策划编辑：马　岚
责任编辑：马　岚
印　　刷：北京京师印务有限公司
装　　订：
出版发行：电子工业出版社
　　　　　北京市海淀区万寿路173信箱　邮编：100036
开　　本：787 × 980　1/16　　印张：44　　字数：1281千字
印　　次：2012年8月第1次印刷
定　　价：89.00元

凡所购买电子工业出版社的图书有缺损问题，请向购买书店调换；若书店售缺，请与本社发行部联系。联系及邮购电话：（010）88254888。
质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。
服务热线：（010）88258888。



Contents

登录华信教育资源网（http://www.hxedu.com.cn）可下载第10章和第11章。
采用本书作为教材的教师可获得本书配套教辅，请联系010-88254555或发邮件至te_services@phei.com.cn



Preface

While significant changes have been made in the current edition from its predecessor, the authors have tried to keep the discussion at the same level of accessibly, that is, less mathematical than the measure theory approach but more rigorous than formula and recipe manuals.

It has been said that *probability* is hard to understand, not so much because of its mathematical underpinnings but because it produces many results that are counter intuitive. Among practically oriented students, *Probability* has many critics. Foremost among these are the ones who ask, "What do we need it for?" This criticism is easy to answer because future engineers and scientists will come to realize that almost every human endeavor involves making decisions in an uncertain or probabilistic environment. This is true for entire fields such as insurance, meteorology, urban planning, pharmaceuticals, and many more. Another, possibly more potent, criticism is, "What good is probability if the answers it furnishes are not certainties but just inferences and likelihoods?" The answer here is that an immense amount of good planning and accurate predictions can be done even in the realm of uncertainty. Moreover, applied probability—often called *statistics*—does provide near certainties: witness the enormous success of political polling and prediction.

In previous editions, we have treaded lightly in the area of statistics and more heavily in the area of random processes and signal processing. In the electronic version of this book, graduate-level signal processing and advanced discussions of random processes are retained, along with new material on statistics. In the hard copy version of the book, we have dropped the chapters on applications to statistical signal processing and advanced topics in random processes, as well as some introductory material on pattern recognition.

The present edition makes a greater effort to reach students with more expository examples and more detailed discussion. We have minimized the use of phrases such as,

"it is easy to show...", "it can be shown...", "it is easy to see...," and the like. Also, we have tried to furnish examples from real-world issues such as the efficacy of drugs, the likelihood of contagion, and the odds of winning at gambling, as well as from digital communications, networks, and signals.

The other major change is the addition of two chapters on elementary statistics and its applications to real-world problems. The first of these deals with parameter estimation and the second with hypothesis testing. Many activities in engineering involve estimating parameters, for example, from estimating the strength of a new concrete formula to estimating the amount of signal traffic between computers. Likewise many engineering activities involve making decisions in random environments, from deciding whether new drugs are effective to deciding the effectiveness of new teaching methods. The origin and applications of standard statistical tools such as the t-test, the Chi-square test, and the F-test are presented and discussed with detailed examples and end-of-chapter problems.

Finally, many self-test multiple-choice exams are now available for students at the book Web site. These exams were administered to senior undergraduate and graduate students at the Illinois Institute of Technology during the tenure of one of the authors who taught there from 1988 to 2006. The Web site also includes an extensive set of small MATLAB programs that illustrate the concepts of probability.

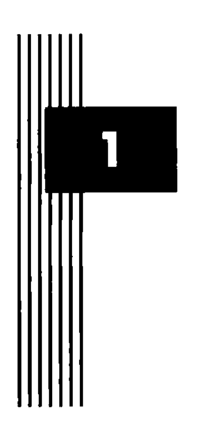
In summary then, readers familiar with the 3rd edition will see the following significant changes:

- A new chapter on a branch of statistics called *parameter estimation* with many illustrative examples;
- A new chapter on a branch of statistics called *hypothesis testing* with many illustrative examples;
- A large number of new homework problems of varying degrees of difficulty to test the student's mastery of the principles of statistics;
- A large number of self-test, multiple-choice, exam questions calibrated to the material in various chapters available on the Companion Web site.
- Many additional illustrative examples drawn from real-world situations where the principles of probability and statistics have useful applications;
- A greater involvement of computers as teaching/learning aids such as (i) graphical displays of probabilistic phenomena; (ii) MATLAB programs to illustrate probabilistic concepts; (iii) homework problems requiring the use of MATLAB/ Excel to realize probability and statistical theory;
- Numerous revised discussions—based on student feedback—meant to facilitate the understanding of difficult concepts.

Henry Stark, IIT
Professor Emeritus

John W. Woods, Rensselaer
Professor

The publishers would like to thank Dr Murari Mitra and Dr Tamaghna Acharya of Bengal Engineering and Science University for reviewing content for the International Edition.



1 Introduction to Probability

1.1 INTRODUCTION: WHY STUDY PROBABILITY?

One of the most frequent questions posed by beginning students of probability is, "Is anything truly random and if so how does one differentiate between the truly random and that which, because of a lack of information, is treated as random but really isn't?" First, regarding the question of truly random phenomena, "Do such things exist?" As we look with telescopes out into the universe, we see vast arrays of galaxies, stars, and planets in apparently random order and position.

At the other extreme from the cosmic scale is what happens at the atomic level. Our friends the physicists speak of such things as the *probability* of an atomic system being in a certain state. The uncertainty principle says that, try as we might, there is a limit to the accuracy with which the position and momentum can be simultaneously ascribed to a particle. Both quantities are fuzzy and indeterminate.

Many, including some of our most famous physicists, believe in an essential randomness of nature. Eugen Merzbacher in his well-known textbook on quantum mechanics [1-1] writes,

> The probability doctrine of quantum mechanics asserts that the indetermination, of which we have just given an example, is a property inherent in nature and not merely a profession of our temporary ignorance from which we expect to be relieved by a future better and more complete theory. The conventional interpretation thus denies the possibility of an ideal theory which would encompass the present quantum mechanics

but would be free of its supposed defects, the most notorious "imperfection" of quantum mechanics being the abandonment of strict classical determinism.

But the issue of determinism versus inherent indeterminism need never even be considered when discussing the validity of the probabilistic approach. The fact remains that there is, quite literally, a nearly uncountable number of situations where we cannot make any categorical deterministic assertion regarding a phenomenon because we cannot measure all the contributing elements. Take, for example, predicting the value of the noise current $i(t)$ produced by a thermally excited resistor R. Conceivably, we might accurately predict $i(t)$ at some instant t in the future if we could keep track, say, of the 10^{23} or so excited electrons moving in each other's magnetic fields and setting up local field pulses that eventually all contribute to producing $i(t)$. Such a calculation is quite inconceivable, however, and therefore we use a probabilistic model rather than Maxwell's equations to deal with resistor noise. Similar arguments can be made for predicting the weather, the outcome of tossing a real physical coin, the time to failure of a computer, dark current in a CMOS imager, and many other situations. Thus, we conclude: Regardless of which position one takes, that is, determinism versus indeterminism, we are forced to use probabilistic models in the real world because we do not know, cannot calculate, or cannot measure all the forces contributing to an effect. The forces may be too complicated, too numerous, or too faint.

Probability is a mathematical model to help us study physical systems in an *average sense*. We have to be able to repeat the experiment many times under the same conditions. Probability then tells us how often to expect the various outcomes. Thus, we cannot use probability in any meaningful sense to answer questions such as "What is the probability that a comet will strike the earth tomorrow?" or "What is the probability that there is life on other planets?" The problem here is that we have no data from similar "experiments" in the past.

R. A. Fisher and R. Von Mises, in the first third of the twentieth century, were largely responsible for developing the groundwork of modern probability theory. The modern axiomatic treatment upon which this book is based is largely the result of the work by Andrei N. Kolmogorov [1-2].

1.2 THE DIFFERENT KINDS OF PROBABILITY

There are essentially four kinds of probability. We briefly discuss them here.

Probability as Intuition

This kind of probability deals with judgments based on intuition. Thus, "She will probably marry him" and "He probably drove too fast" are in this category. Intuitive probability can lead to contradictory behavior. Joe is still likely to buy an imported Itsibitsi, world famous for its reliability, even though his neighbor Frank has a 19-year-old Buick that has never broken down and Joe's other neighbor, Bill, has his Itsibitsi in the repair shop. Here Joe may be behaving "rationally," going by the statistics and ignoring, so-to-speak, his personal observation. On the other hand, Joe will be wary about letting his nine-year-old

daughter Jane swim in the local pond, if Frank reports that Bill thought that he might have seen an alligator in it. This despite the fact that no one has ever reported seeing an alligator in this pond, and countless people have enjoyed swimming in it without ever having been bitten by an alligator. To give this example some credibility, assume that the pond is in Florida. Here Joe is ignoring the statistics and reacting to, what is essentially, a rumor. Why? Possibly because the *cost* to Joe "just-in-case" there is an alligator in the pond would be too high [1-3].

People buying lottery tickets intuitively believe that certain number combinations like month/day/year of their grandson's birthday are more likely to win than say, 06–06–06. How many people will bet even odds that a coin that, heretofore has behaved "fairly," that is, in an unbiased fashion, will come up heads on the next toss, *if in the last seven tosses it has come up heads?* Many of us share the belief that the coin has some sort of memory and that, after seven heads, that coin must "make things right" by coming up with more tails.

A mathematical theory dealing with intuitive probability was developed by B. O. Koopman [1-4]. However, we shall not discuss this subject in this book.

Probability as the Ratio of Favorable to Total Outcomes (Classical Theory)

In this approach, which is not experimental, the probability of an event is computed *a priori*[†] by counting the number of ways n_E that E can occur and forming the ratio n_E/n, where n is the number of all possible outcomes, that is, the number of all alternatives to E plus n_E. An important notion here is that all outcomes are equally likely. Since equally likely is really a way of saying equally probable, the reasoning is somewhat circular. Suppose we throw a pair of unbiased six-sided dice[‡] and ask what is the probability of getting a 7. We partition the outcome space into 36 equally likely outcomes as shown in Table 1.2-1, where each entry is the sum of the numbers on the two dice.

Table 1.2-1 Outcomes of Throwing Two Dice

	1st die					
2nd die	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

[†] *A priori* means relating to reasoning from self-evident propositions or prior experience. The related phrase, *a posteriori* means relating to reasoning from observed facts.

[‡] We will always assume that our dice have six sides.

The total number of outcomes is 36 if we keep the dice distinct. The number of ways of getting a 7 is $n_7 = 6$. Hence

$$P[\text{getting a } 7] = \frac{6}{36} = \frac{1}{6}.$$

Example 1.2-1

(*toss a fair coin twice*) The possible outcomes are HH, HT, TH, and TT. The probability of getting at least one tail T is computed as follows: With E denoting the event of getting at least one tail, the event E is the set of outcomes

$$E = \{\text{HT}, \text{TH}, \text{TT}\}.$$

Thus, event E occurs whenever the outcome is HT or TH or TT. The number of elements in E is $n_E = 3$; the number of all outcomes N, is four. Hence

$$P[\text{at least one T}] = \frac{n_E}{n} = \frac{3}{4}.$$

Note that since no physical experimentation is involved, there is no problem in postulating an ideal "fair coin." Effectively, in classical probability every experiment is considered "fair."

The classical theory suffers from at least two significant problems: (1) It cannot deal with outcomes that are not equally likely; and (2) it cannot handle an infinite number of outcomes, that is when $n = \infty$. Nevertheless, in those problems where it is impractical to actually determine the outcome probabilities by experimentation and where, because of symmetry considerations, one can indeed argue equally likely outcomes, the classical theory is useful.

Historically, the classical approach was the predecessor of Richard Von Mises' [1-6] relative frequency approach developed in the 1930s, which we consider next.

Probability as a Measure of Frequency of Occurrence

The relative frequency approach to defining the probability of an event E is to perform an experiment n times. The number of times that E appears is denoted by n_E. Then it is tempting to define the probability of E occurring by

$$P[E] = \lim_{n\to\infty} \frac{n_E}{n}. \tag{1.2-1}$$

Quite clearly since $n_E \leq n$ we must have $0 \leq P[E] \leq 1$. One difficulty with this approach is that we can never perform the experiment an infinite number of times, so we can only estimate $P[E]$ from a finite number of trials. Secondly, we *postulate* that n_E/n approaches a limit as n goes to infinity. But consider flipping a fair coin 1000 times. The likelihood of getting exactly 500 heads is very small; in fact, if we flipped the coin 10,000 times, the likelihood of getting exactly 5000 heads is even smaller. As $n \to \infty$, the event of observing

exactly $n/2$ heads becomes vanishingly small. Yet our intuition demands that $P[\text{head}] = \frac{1}{2}$ for a fair coin. Suppose we choose a $\delta > 0$; then we shall find experimentally that if the coin is truly fair, the number of times that

$$\left|\frac{n_E}{n} - \frac{1}{2}\right| > \delta, \tag{1.2-2}$$

as n becomes large, becomes very small. Thus, although it is very unlikely that at any stage of this experiment, especially when n is large, n_E/n is exactly $\frac{1}{2}$, this ratio will nevertheless hover around $\frac{1}{2}$, and the number of times it will make significant excursion away from the vicinity of $\frac{1}{2}$ according to Equation 1.2-2, becomes very small indeed.

Despite these problems with the relative frequency definition of probability, the relative frequency concept is essential in applying probability theory to the physical world.

Example 1.2-2

(*random.org*) An Internet source of random numbers is RANDOM.ORG, which was founded by a professor in the School of Computer Science and Statistics at Trinity College, Dublin, Ireland. It calculates random digits as a function of atmospheric noise and has passed many statistical tests for true randomness. Using one of the site's free services, we have downloaded 10,000 random numbers, each taking on values from 1 to 100 equally likely. We have written the MATLAB function `RelativeFrequencies()` that takes this file of random numbers and plots the ratio n_E/n as a function of the trial number $n = 1, \ldots, 10,000$. We can choose the event E to be the occurrence of any one of the 100 numbers. For example for $E \triangleq \{\text{occurrence of number 5}\}$, the number n_E counts the number of times 5 has occurred among the 10,000 numbers up to position n. A resulting output plot is shown in Figure 1.2-1, where we see a general tendency toward convergence to the ideal value of $0.01 = 1/100$ for 100 equally likely numbers. An output plot for another number choice 23 is shown in Figure 1.2-2 again showing a general tendency to converge to the ideal value here of 0.01. In both cases though, we note that the convergence is not exact at any value of n, but rather just convergence to a small neighborhood of the ideal value.

This program is available at this book's website.

Probability Based on an Axiomatic Theory

The axiomatic approach is followed in most modern textbooks on the subject. To develop it we must introduce certain ideas, especially those of a random experiment, a sample space, and an event. Briefly stated, a random experiment is simply an experiment in which the outcomes are nondeterministic, that is, more than one outcome can occur each time the experiment is run. Hence the word *random* in *random experiment*. The *sample space* is the set of all outcomes of the random experiment. An *event* is a subset of the sample space that satisfies certain constraints. For example, we want to be able to calculate the probability for each event. Also in the case of noncountable or continuous sample spaces, there are certain technical restrictions on what subsets can be called events. An event with only one outcome will be called a *singleton* or *elementary* event. These notions will be made more precise in Sections 1.4 and 1.5.

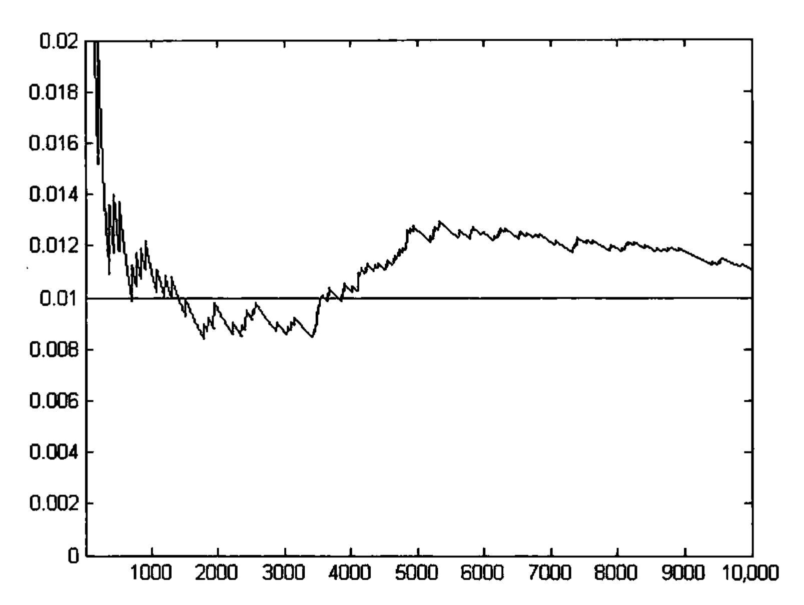


Figure 1.2-1 Plot of n_E/n for $E = \{\text{occurrence of number 5}\}$ versus n from atmospheric noise (from website RANDOM.ORG).

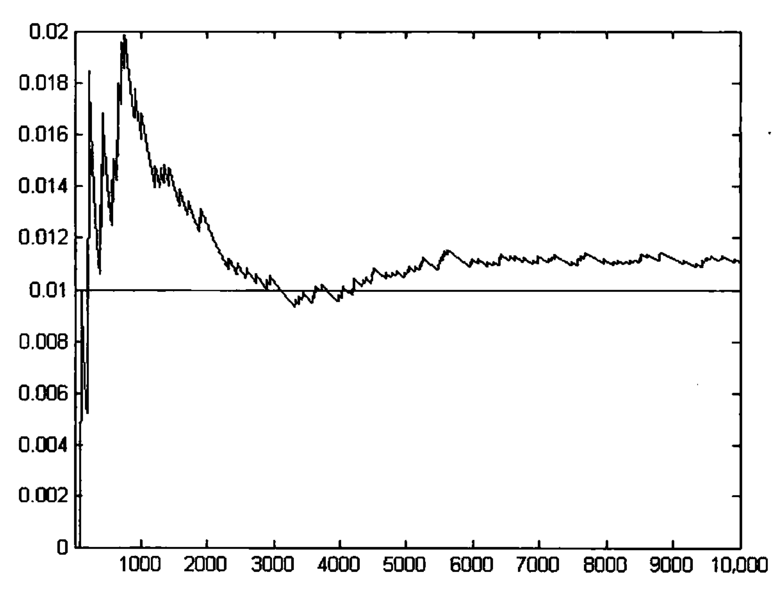


Figure 1.2-2 Plot of n_E/n for $E = \{\text{occurrence of number 23}\}$ versus n from atmospheric noise (from website RANDOM.ORG).

1.3 MISUSES, MISCALCULATIONS, AND PARADOXES IN PROBABILITY

The misuse of probability and statistics in everyday life is quite common. Many of the misuses are illustrated by the following examples. Consider a defendant in a murder trial who pleads not guilty to murdering his wife. The defendant has on numerous occasions beaten his wife. His lawyer argues that, yes, the defendant has beaten his wife but that among men who do so, the probability that one of them will actually murder his wife is only 0.001, that is, only one in a thousand. Let us assume that this statement is true. It is meant to sway the jury by implying that the fact of beating one's wife is no indicator of murdering one's wife. Unfortunately, unless the members of the jury have taken a good course in probability, they might not be aware that a far more significant question is the following: *Given that a battered wife is murdered, what is the probability that the husband is the murderer?* Statistics show that this probability is, in fact, *greater than one-half.*

In the 1996 presidential race, Senator Bob Dole's age became an issue. His opponents claimed that a 72-year-old white male has a 27 percent risk of dying in the next five years. Thus it was argued, were Bob Dole elected, the probability that he would fail to survive his term was greater than one-in-four. The trouble with this argument is that the probability of survival, as computed, was not conditioned on additional pertinent facts. As it happens, if a 72-year-old male is still in the workforce and, additionally, happens to be rich, then taking these additional facts into consideration, the average 73-year-old (the age at which Dole would have assumed the presidency) *has only a one-in-eight chance of dying in the next four years* [1-3].

Misuse of probability appears frequently in predicting life elsewhere in the universe. In his book *Probability 1* (Harcourt Brace & Company, 1998), Amir Aczel assures us that we can be certain that alien life forms are out there just waiting to be discovered. However, in a cogent review of Aczel's book, John Durant of London's Imperial College writes,

> Statistics are extremely powerful and important, and Aczel is a very clear and capable exponent of them. But statistics cannot substitute for empirical knowledge about the way the universe behaves. We now have no plausible way of arriving at robust estimates about the way the universe behaves. We now have no plausible way of arriving at robust estimates for the probability of life arriving spontaneously when the conditions are right. So, until we either discover extraterrestrial life or understand far more about how at least one form of life—terrestrial life—first appeared, we can do little more than guess at the likelihood that life exists elsewhere in the universe. And as long as we're guessing, we should not dress up our interesting speculations as mathematical certainties.

The computation of probabilities based on relative frequency can lead to paradoxes. An excellent example is found in [1-3]. We repeat the example here:

> In a sample of American women between the ages of 35 and 50, 4 out of 100 develop breast cancer within a year. Does Mrs. Smith, a 49-year-old American woman, therefore have a 4% chance of getting breast cancer in the next year? There is no answer. Suppose that in a sample of women between the ages of 45 and 90—a class to which

Mrs. Smith also belongs—11 out of 100 develop breast cancer in a year. Are Mrs. Smith's chances 4%, or are they 11%? Suppose that her mother had breast cancer, and 22 out of 100 women between 45 and 90 whose mothers had the disease will develop it. Are her chances 4%, 11%, or 22%? She also smokes, lives in California, had two children before the age of 25 and one after 40, is of Greek descent .... What group should we compare her with to figure out the "true" odds? You might think, the more specific the class, the better—but the more specific the class, the smaller its size and the less reliable the frequency. If there were only two people in the world very much like Mrs. Smith, and one developed breast cancer, would anyone say that Mrs. Smith's chances are 50%? In the limit, the only class that is truly comparable with Mrs. Smith in all her details is the class containing Mrs. Smith herself. But in a class of one "relative frequency" makes no sense.

The previous example should not leave the impression that the study of probability, based on relative frequency, is useless. For one, there are a huge number of engineering and scientific situations that are not nearly as complex as the case of Mrs. Smith's likelihood of getting cancer. Also, it is true that if we refine the class and thereby reduce the class size, our estimate of probability based on relative frequency becomes less stable. But exactly how much less stable is deep within the realm of the study of probability and its offspring *statistics* (e.g., see the Law of Large Numbers in Section 4.4). Also, there are many situations where the required conditioning, that is, class refinement, is such that the class size is sufficiently large for excellent estimates of probability. And finally returning to Mrs. Smith, if the class size starts to get too small, then stop adding conditions and learn to live with a probability estimate associated with a larger, less refined class. This estimate may be sufficient for all kinds of actions, that is, planning screening tests, and the like.

1.4 SETS, FIELDS, AND EVENTS

A set is a collection of objects, either concrete or abstract. An example of a set is the set of all New York residents whose height equals or exceeds 6 feet. A subset of a set is a collection that is contained within the larger set. Thus, the set of all New York City residents whose height is between 6 and $6\frac{1}{2}$ feet is a subset of the previous set. In probability theory we call sets *events*. We are particularly interested in the set of all outcomes of a random experiment and subsets of this set. We denote the set of all outcomes by Ω, and individual outcomes by ζ.[†] The set Ω is called the *sample space* of the random experiment. Certain subsets of Ω, whose probabilities we are interested in, are called *events*. In particular Ω itself is called the *certain event* and the empty ϕ set is called the *null event*.

Examples of Sample Spaces

Example 1.4-1

(*coin flip*) The experiment consists of flipping a coin once. Then $\Omega = \{\text{H}, \text{T}\}$, where H is a head and T is a tail.

[†]Greek letter ζ is pronounced zeta.

Example 1.4-2

(*coin flip twice*) The experiment consists of flipping a coin twice. Then $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. One of sixteen subsets of Ω is E=$\{\text{HH}, \text{HT}, \text{TH}\}$; it is the event of getting at least one head in two flips.

Example 1.4-3

(*hair on head*) The experiment consists of choosing a person at random and counting the hairs on his or her head. Then

$$\Omega = \{0, 1, 2, \ldots, 10^7\},$$

that is, the set of all nonnegative integers up to 10^7, it being assumed that no human head has more than 10^7 hairs.

Example 1.4-4

(*couple's ages*) The experiment consists of determining the age to the nearest year of each member of a married couple chosen at random. Then with x denoting the age of the man and y denoting the age of the woman, Ω is described by

$$\Omega = \{\text{2-tuples } (x, y):\ x \text{ any integer in } 10\text{–}200;\ y \text{ any integer in } 10\text{–}200\}.$$

Note that in Example 1.4-4 we have assumed that no human lives beyond 200 years and that no married person is ever less than ten years old. Similarly, in Example 1.4-1, we assumed that the coin never lands on edge. If the latter is a possible outcome, it must be included in Ω in order for it to denote the set of *all* outcomes as well as the certain event.

Example 1.4-5

(*angle in elastic collision*) The experiment consists of observing the angle of deflection of a nuclear particle in an elastic collision. Then

$$\Omega = \{\theta:\ -\pi \le \theta \le \pi\}.$$

An example of an event or subset of Ω is

$$E = \left\{\frac{\pi}{4} \le \theta \le \frac{\pi}{4}\right\} \subset \Omega.$$

Example 1.4-6

(*electrical power*) The experiment consists of measuring the instantaneous power P consumed by a current-driven resistor. Then

$$\Omega = \{P:\ P \ge 0\}.$$

Since power cannot be negative, we leave out negative values of P in Ω. A subset of Ω is the event $E = \{P > 10^{-3} \text{ watts}\}$.

Note that in Examples 1.4-5 and 1.4-6, the number of elements in Ω is uncountably infinite. Therefore, there are an uncountably infinite number of subsets. When, as in Example 1.4-4, the number of outcomes is finite, the number of distinct subsets is also finite, and each represents an event. Thus, if $\Omega = \{\zeta_1, \ldots, \zeta_N\}$, the number of possible subsets of Ω is 2^N. We can see this by noting that each element ζ_i either is or is not present in any given subset of Ω. This gives rise to 2^N distinct subsets or events, including the certain event and the impossible or null event.

Review of set theory. The *union* (sum) of two sets E and F, written $E \cup F$ or $E + F$, is the set of all elements that are in at least one of the sets E and F. Thus, with $E = \{1, 2, 3, 4\}$ and $F = \{1, 3, 4, 5, 6\}$,†

$$E \cup F = \{1, 2, 3, 4, 5, 6\}.$$

If E is a subset of F, we indicate this by writing $E \subset F$. Clearly for $E \subset F$ it follows that $E \cup F = F$. We indicate that ζ is an element of Ω or "belongs" to Ω by writing $\zeta \in \Omega$. Thus, we can write

$$E \cup F = \{\zeta \colon \zeta \in E \text{ or } \zeta \in F\}, \tag{1.4-1}$$

where the "or" here is inclusive. Clearly $E \cup F = F \cup E$. The *intersection* or set product of two sets E and F, written $E \cap F$ or just EF, is the set of elements common to both E and F. Thus, in the preceding example

$$EF = \{1, 3, 4\}.$$

Formally, $EF \triangleq \{\zeta \colon \zeta \in E \text{and } \zeta \in F\} = FE$. The *complement* of a set E, written E^c, is the set of *all elements not in* E. From this it follows that if Ω is the sample space or, more generally, the universal set, then

$$E \cup E^c = \Omega. \tag{1.4-2}$$

Also $EE^c = \phi$. The *set difference* of two sets or, more appropriately, the *reduction* of E by F, written $E - F$, is the set made up of elements in E that are not in F. It should be clear that

$$E - F \triangleq EF^c,$$

$$F - E \triangleq FE^c,$$

but be careful. Set difference does not behave like difference of numbers, for example, $F - E - E = F - E$. The *exclusive or* of two sets, written $E \oplus F$, is the set of all elements in E *or* F but not both. It is readily shown that‡

$$E \oplus F = (E - F) \cup (F - E). \tag{1.4-3}$$

†Remember, the order of the elements in a set is not important.

‡Equation 1.4-3 shows why $\cup$ is preferable to $+$ to indicate union. The beginning student might—in error—write $(E - F) + (F - E) = E - F + F - E = 0$, which is meaningless. Note also that $F + F \neq 2F$, which is also meaningless. In fact $F + F = F$. So, only use $+$ and $-$ operators in set theory with care.

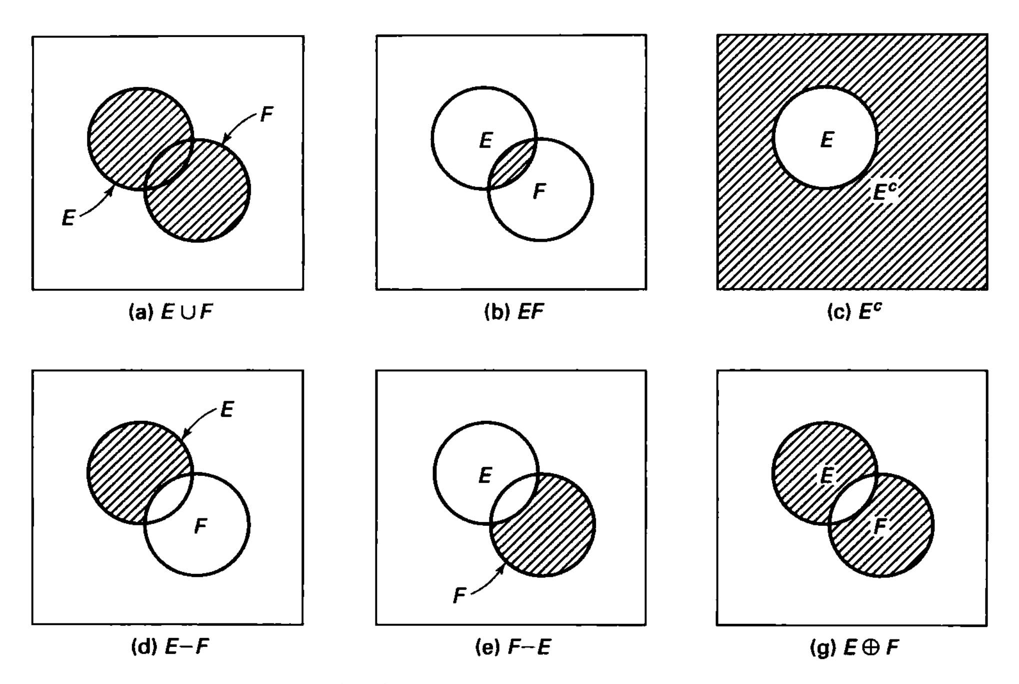


Figure 1.4-1 Venn diagrams for set operations.

The operation of unions, intersections, and so forth can be illustrated by Venn diagrams, which are useful as aids in reasoning and in establishing probability relations. The various set operations $E \cup F$, EF, E^c, $E - F$, $F - E$, $E \oplus F$ are shown in Figure 1.4-1 in hatch lines.

Two sets E, F are said to be disjoint if $EF = \phi$; that is, they have no elements in common. Given any set E, an *n-partition* of E consists of a sequence of sets E_i, where $i = 1, \ldots, n$, such that $E_i \subset E$, $\bigcup_{i=1}^{n} E_i = E$, and $E_i E_j = \phi$ for all $i \neq j$. Thus, given two sets E, F, a 2-partition of F is

$$F = FE \cup FE^c. \tag{1.4-4}$$

It is easy to see, using Venn diagrams, the following results:

$$(E \cup F)^c = E^c F^c \tag{1.4-5}$$

$$(EF)^c = E^c \cup F^c \tag{1.4-6}$$

and, by induction,† given sets $E_1, \ldots, E_n$:

†See Section A.4 in Appendix A for the meaning of *mathematical induction*.

$$\left[\bigcup_{i=1}^{n} E_i\right]^c = \bigcap_{i=1}^{n} E_i^c \tag{1.4-7}$$

$$\left[\bigcap_{i=1}^{n} E_i\right]^c = \bigcup_{i=1}^{n} E_i^c. \tag{1.4-8}$$

The relations are known as De Morgan's laws after the English mathematician Augustus De Morgan (1806–1871).

While Venn diagrams allow us to visualize these results, they don not really achieve their proof. Towards this end, consider the mathematical definition of equality of two sets. Two sets E and F are said to be equal if every element in E is in F and vice versa. Equivalently,

$$E = F \quad \text{if} \quad E \subset F, \text{ and } F \subset E. \tag{1.4-9}$$

Example 1.4-7

(*proving equality of sets*) If we want to strictly prove one of the above set equalities, say Eq. 1.4-4, $F = FE \cup FE^c$, we must proceed as follows. First show $F \subset FE \cup FE^c$ and then show $F \supset FE \cup FE^c$. To show $F \subset FE \cup FE^c$, we consider an arbitrary element $\varsigma \in F$, then ς must be in either FE or FE^c for any set E, and thus ς must be in $FE \cup FE^c$. This establishes that $F \subset FE \cup FE^c$. Going the other way, to show $F \supset FE \cup FE^c$, we start with an arbitrary element $\varsigma \in FE \cup FE^c$. It must be that ς belongs to either FE or FE^c and so ς must belong to F, thus establishing $F \supset FE \cup FE^c$. Since we have shown the two set inclusions, we may write $F = FE \cup FE^c$ meaning that both sets are equal.

Using this method you can establish the following helpful laws of set theory:

1. *associative law for union*

$$A \cup (B \cup C) = (A \cup B) \cup C$$

2. *associative law for intersection*

$$A(BC) = (AB)C$$

3. *distributive law for union*

$$A \cup (BC) = (A \cup B)(A \cup C)$$

4. *distributive law for intersection*

$$A(B \cup C) = (AB) \cup (AC)$$

We will use these identities or laws for analyzing set equations below. However, these four laws must be proven first. Here, as an example, we give the proof of 1.

Example 1.4-8

(*proof of associative law for union*) We wish to prove $A \cup (B \cup C) = (A \cup B) \cup C$. To do this we must show that both $A \cup (B \cup C) \subseteq (A \cup B) \cup C$ and $A \cup (B \cup C) \supseteq (A \cup B) \cup C$. Starting with the former, assume that $\zeta \in A \cup (B \cup C)$; then it follows that ζ is in A or in $B \cup C$. But then ζ is in A or B or C, so it is in $A \cup B$ or in C, which is the same as saying $\zeta \in (A \cup B) \cup C$. To complete the proof, we must go the other way and show starting with $\zeta \in (A \cup B) \cup C$ that ζ must also be an element of $A \cup (B \cup C)$. This part is left for the student.

Sigma fields. Consider a universal set Ω and a certain collection of subsets of Ω. Let E and F be two arbitrary subsets in this collection. This collection of subsets forms a *field* $\mathscr{M}$ if

(1) $\phi \in \mathscr{M}$, $\Omega \in \mathscr{M}$.
(2) If $E \in \mathscr{M}$ and $F \in \mathscr{M}$, then $E \cup F \in \mathscr{M}$, and $EF \in \mathscr{M}$.†
(3) If $E \in \mathscr{M}$, then $E^c \in \mathscr{M}$.

We will need to consider fields of sets (fields of events in probability) in order to avoid some problems. If our collection of events were not a field, then we could define a probability for some events, but not for their complement; that is, we could not define the probability that these events do not occur! Similarly we need to be able to consider the probability of the union of any two events, that is, the probability that either 'or both' of the two events occur. Thus, for probability theory, we need to have a probability assigned to all the events in the field.

Many times we will have to consider an infinite set of outcomes and events. In that case we need to extend our definition of field. A *sigma* (σ) *field*‡ $\mathscr{F}$ is a field that is *closed* under any countable number of unions, intersections, and complementations. Thus, if $E_1, \ldots, E_n, \ldots$ belong to $\mathscr{F}$ so do

$$\bigcup_{i=1}^{\infty} E_i \quad \text{and} \quad \bigcap_{i=1}^{\infty} E_i,$$

where these are simply defined as

$$\bigcup_{i=1}^{\infty} E_i \triangleq \{\text{the set of all elements in } \textit{at least one } E_i\}$$

and

$$\bigcap_{i=1}^{\infty} E_i \triangleq \{\text{the set of all elements in } \textit{every } E_i\}.$$

Note that these two infinite operations of union and intersection would be meaningless without a specific definition. Unlike infinite summations which are defined by limiting operations with numbers, there are no such limiting operations defined on sets, hence the need for a definition.

Events. Consider a probability experiment with sample space Ω. If Ω has a countable number of elements, then every subset of Ω may be assigned a probability in a way consistent with the axioms given in the next section. Then the class of all subsets will make up a field or σ-field simply because every subset is included. This collection of all the subsets of Ω is called the *largest* σ-field.

---

†From this it follows by mathematical induction, that if $E_1, \ldots, E_n$ belongs to $\mathscr{M}$ so do $\bigcup_{i=1}^{n} E_i \in \mathscr{M}$ and $\bigcap_{i=1}^{N} E_i \in \mathscr{M}$.

‡Also sometimes called a σ-algebra.

Sometimes though we do not have enough probability information to assign a probability to every subset. In that case we need to define a smaller field of events that is still a σ-field, but just a smaller one. We will discuss this matter in the example below. Going to the limit, if we have *no* probability information, then we must content ourselves with the *smallest* σ-field of events consisting of just the null event ϕ and the certain event Ω. While this collection of two events is a σ-field it is not very useful.

Example 1.4-9

(*field generated by two events*) Assume we have interest in only two events A and B in an arbitrary sample space Ω and we desire to find the smallest field containing these two events. We can proceed as follows. First we generate a disjoint decomposition of the sample space as follows.

$$\begin{aligned}\Omega &= \Omega(A \cup A^c)(B \cup B^c) \\ &= AB \cup AB^c \cup A^cB \cup A^cB^c.\end{aligned}$$

Next generate a collection of events from these four basic disjoint (non-overlapping) events as follows: The first four events are AB, AB^c A^cB, and A^cB^c. Then we add the pairwise unions of these disjoint events: $AB \cup AB^c$, $AB \cup A^cB$, and $AB \cup A^cB^c$. Finally we add the unions of tripples of these four disjoint events. The total number of events will then be $2 \times 2 \times 2 \times 2 = 2^4 = 16$, since each of the four basic disjoint events can be included, or not, in the event.

This collection of events is guarenteed to be a field, since we construct each of its 16 events from the four basic disjoint events, thus ensuring that complements are in the collection via $\Omega = AB \cup AB^c \cup A^cB \cup A^cB^c$. Unions are trivially in the collection too. Because all the events in the collection are built up from the four disjoint events, complements are just the events that have been left out, eg. $(AB \cup AB^c)^c = A^cB \cup A^cB^c$ which is recognized as being in the collection. Hence we have a field. In fact this is the smallest field that contains the events A and B. We call this the field generated by events A and B. Can you show that event A is in this field?

When Ω is not countable, for example, when $\Omega = R^1$ = the real line, advanced mathematics (*measure theory*) has found that not every subset of Ω can be assigned a probability (is measurable) in a way that will be consistent. So we must content ourselves with smaller collections of subsets of the universal event Ω that form a σ-field. On the real line R^1 for example, we can generate a σ-field from all the intervals, open/closed, and this is called the *Borel field of events* on the real line. As a practical matter, it has been found that the Borel field on the real line includes all subsets of engineering and scientific interest.†

At this stage of our development, we have two of the three objects required for the axiomatic theory of probability, namely, a sample space Ω of outcomes ζ, and a σ-field $\mathscr{F}$ of events defined on Ω. We still need a probability measure P. The three objects $(\Omega, \mathscr{F}, P)$ form a triple called the *probability space* $\mathscr{P}$ that will constitute our mathematical model. However, the probability measure P must satisfy the following three axioms due to Kolmogorov.

†For two-dimensional Euclidean sample spaces, the Borel field of events would be subsets of $R^1 \times R^1 = R^2$; for three-dimensional sample spaces, it would be subsets of $R^1 \times R^1 \times R^1 = R^3$.

1.5 AXIOMATIC DEFINITION OF PROBABILITY

Probability is a set function $P[\cdot]$ that assigns to every event $E \in \mathscr{F}$ a number $P[E]$ called the probability of event E such that

(1) $P[E] \geq 0$. (1.5-1)
(2) $P[\Omega] = 1$. (1.5-2)
(3) $P[E \cup F] = P[E] + P[F] \quad \text{if} \quad EF = \phi$. (1.5-3)

The probability measure is not like an ordinary function. It does not take numbers for its argument, but rather it takes sets; that is, it is a measure of sets, our mathematical model for events. Since this is a special function, to distinguish it we will always use square brackets for its argument, a set of outcomes ζ in the sample space Ω.

These three axioms are sufficient† to establish the following basic results, all but one of which we leave as exercises for the reader. Let E and F be events contained in $\mathscr{F}$, then

(4) $P[\phi] = 0$. (1.5-4)
(5) $P[EF^c] = P[E] - P[EF]$, (1.5-5)
(6) $P[E] = 1 - P[E^c]$. (1.5-6)
(7) $P[E \cup F] = P[E] + P[F] - P[EF]$. (1.5-7)

From Axiom 3 we can establish by mathematical induction that

$$P\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} P[E_i] \text{ if } E_iE_j = \phi \quad \text{for all} \quad i \neq j. \tag{1.5-8}$$

From this result and Equation 1.5-7, we can establish by induction, the general result that $P\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} P[E_i]$. This result is sometimes known as the *union bound*, often used in digital communications theory to provide an upper bound on the probability of error.

Example 1.5-1

(*probability of the union of two events*) We wish to prove result (7). First we decompose the event $E \cup F$ into three disjoint events as follows:

$$E \cup F = EF^c \cup E^cF \cup EF.$$

By Axiom 3

$$\begin{aligned} P[E \cup F] &= P[EF^c \cup E^cF] + P[EF] \\ &= P[EF^c] + P[E^cF] + P[EF], \text{ by Axiom 3 again} \\ &= P[E] - P[EF] + P[F] - P[EF] + P[EF] \\ &= P[E] + P[F] - P[EF]. \end{aligned} \tag{1.5-9}$$

† A fourth axiom: $P\left[\bigcup_{i=1}^{\infty} E_i\right] = \sum_{i=1}^{\infty} P[E_i]$ if $E_iE_j = \phi$ for all $i \neq j$ must be included to enable one to deal rigorously with limits and countable unions. This axiom is of no concern to us here but will be in later chapters.

We can apply this result to the following problem.

In a certain bread store, there are two events of interest $W \triangleq \{\text{white bread is available}\}$ and $R \triangleq \{\text{rye bread is available}\}$. Based on past experience with this establishment, we take $P[W] = 0.8$ and, $P[R] = 0.7$. We also know that the probability that both breads are present is 0.6, that is, $P[WR] = 0.6$. We now ask what is the probability of either bread being present, that is, what is $P[W \cup R]$? The answer is obtained basic result (7) as

$$\begin{aligned} P[W \cup R] &= P[W] + P[R] - P[WR] \\ &= 0.8 + 0.7 - 0.6 \\ &= 0.9. \end{aligned}$$

We pause for a bit of terminology. We say an event E *occurs* whenever the outcome of our experiment is one of the elements in E. So "$P[E]$" is read as "the probability that event E occurs."

A measure of events not outcomes. The reader will have noticed that we talk of the probability of events rather than the probability of outcomes. For finite and countable sample spaces, we could just as well talk about probabilities of outcomes; however, we do not do so for several reasons. One is we would still need to talk about probabilities of events and so would need two types of probability measures, one for outcomes and one for events. Second, in some cases we only know the probability of some events and don't know the probabilistic detail to assign a probability to each outcome. Lastly, and most importantly, in the case of continuous sample spaces with uncountable outcomes, for example the points on the real number interval $[0, 5]$ these may well have zero probability, and hence any theory based on probability of the outcomes would be useless. For these and other reasons we base our approach on events, and so probability measures events not outcomes.

Example 1.5-2

(*toss coin once*) The experiment consists of throwing a coin once. Our idealized outcomes are then H and T, with sample space:

$$\Omega = \{\text{H}, \text{T}\}.$$

The σ-field of events consists of the following sets: $\{\text{H}\}$, $\{\text{T}\}$, Ω, ϕ. With the coin assumed fair, we have†

$$P[\{\text{H}\}] = P[\{\text{T}\}] = \tfrac{1}{2}, \qquad P[\Omega] = 1, \qquad P[\phi] = 0.$$

Example 1.5-3

(*toss die once*) The experiment consists of throwing a die once. The outcomes are the number of dots $\varsigma = 1, \ldots, 6$, appearing on the upward facing side of the die. The sample

---

†Remember the outcome ζ is the output or result of our experiment. The set of all outcomes is the sample space Ω.

space Ω is given by $\Omega = \{1,2,3,4,5,6\}$. The event field consists of 2^6 events, each one containing, or not, each of the outcomes i. Some events are

$$\phi, \Omega, \{1\}, \{1,2\}, \{1,2,3\}, \{1,4,6\}, \text{ and } \{1,2,4,5\}.$$

We assign probabilities to the elementary or singleton events $\{\zeta\}$:

$$P[\{\zeta\}] = \tfrac{1}{6} \qquad i = 1, \ldots, 6.$$

All probabilities can now be computed from the basic axioms and the assumed probabilities for the elementary events. For example, with $A = \{1\}$ and $B = \{2,3\}$ we obtain $P[A] = \frac{1}{6}$. Also $P[A \cup B] = P[A] + P[B]$, since $AB = \phi$. Furthermore, $P[B] = P[\{2\}] + P[\{3\}] = \frac{2}{6}$ so that

$$P[A \cup B] = \tfrac{1}{6} + \tfrac{2}{6} = \tfrac{1}{2}.$$

Example 1.5-4

(*choose ball from urn*) The experiment consists of picking at random a numbered ball from 12 balls numbered 1 to 12 in an urn. Our idealized outcomes are then the numbers $\zeta = 1$ to 12, with sample space:

$$\Omega = \{1, \ldots, 12\}.$$

Let the following events be specified

$$A^{\dagger} = \{1, \ldots, 6\}, \qquad B = \{3, \ldots, 9\}$$

$$A \cup B = \{1, \ldots, 9\}, \qquad AB = \{3,4,5,6\}, \qquad AB^c = \{1,2\}$$

$$B^c = \{1,2,10,11,12\}, \qquad A^c = \{7, \ldots, 12\}, \qquad A^cB^c = \{10,11,12\}$$

$$(AB)^c = \{1,2,7,8,9,10,11,12\}.$$

Hence

$$P[A] = P[\{1\}] + P[\{2\}] + \ldots + P[\{6\}],$$

$$P[B] = P[\{3\}] + \ldots + P[\{9\}],$$

$$P[AB] = P[\{3\}] + \ldots + P[\{6\}].$$

If $P[\{1\}] = \ldots = P[\{12\}] = \frac{1}{12}$, then $P[A] = \frac{1}{2}$, $P[B] = \frac{7}{12}$, $P[AB] = \frac{4}{12}$, and so forth.

We point out that a theory of probability could be developed from a slightly different set of axioms [1-7]. However, whatever axioms are used and whatever theory is developed, for it to be useful in solving problems in the physical world, it must model our empirical concept of probability as a relative frequency and the consequences that follow from it.

---

†We say event A *occurs* whenever any of the number 1 through 6 appears on a ball removed from the urn.

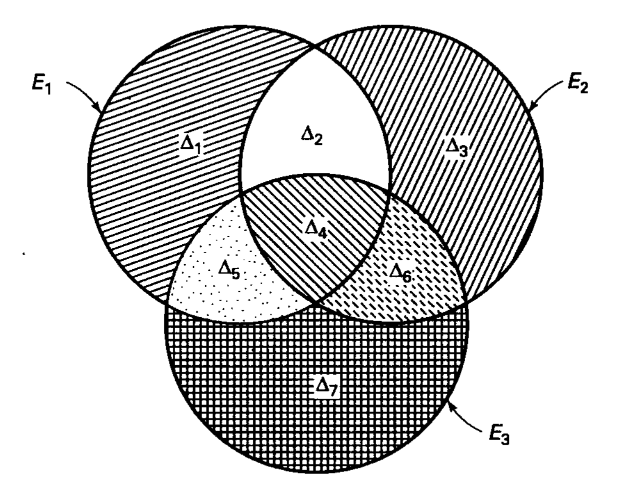


Figure 1.5-1 Partitioning $\bigcup_{i=1}^{3} E_i$ into seven disjoint regions $\Delta_1, \ldots, \Delta_7$.

Probability of union of events. The extension of Equation 1.5-7 to the case of three events is straightforward but somewhat tedious. We consider three events, E_1, E_2, E_3, and wish to compute the probability, $P[\bigcup_{i=1}^{3} E_i]$, that at least one of these events occurs. From the Venn diagram in Figure 1.5-1, we see seven disjoint regions in $\bigcup_{i=1}^{3} E_i$ which we label as Δ_i, $i = 1, \ldots, 7$. You can prove it using the same method used in Example 1.4-9. Then $P[\bigcup_{i=1}^{3} E_i] = P\,[\bigcup_{i=1}^{7} \Delta_i] = \sum_{i=1}^{7} P[\Delta_i]$, from Axiom 3.

In terms of the original events, the seven disjoint regions can be identified as

$$\Delta_1 = E_1 E_2^c E_3^c = E_1(E_2 \cup E_3)^c,$$

$$\Delta_2 = E_1 E_2 E_3^c,$$

$$\Delta_3 = E_1^c E_2 E_3^c = E_2(E_1 \cup E_3)^c,$$

$$\Delta_4 = E_1 E_2 E_3,$$

$$\Delta_5 = E_1 E_2^c E_3,$$

$$\Delta_6 = E_1^c E_2 E_3,$$

$$\Delta_7 = E_1^c E_2^c E_3 = E_3(E_1 \cup E_2)^c.$$

The computations of the probabilities $P[\Delta_i]$, $i = 1, \ldots, 7$, follow from Equations 1.5-5 and 1.5-7. Thus, we compute

$$\begin{aligned} P[\Delta_1] &= P[E_1] - P[E_1E_2 \cup E_1E_3] \\ &= P[E_1] - \{P[E_1E_2] + P[E_1E_3] - P[E_1E_2E_3]\}. \end{aligned}$$

In obtaining the first line, we used Equation 1.5-5. In obtaining the second line, we used Equation 1.5-7. The computations of $P[\Delta_i]$, $i = 3, 7$, are quite similar to the computation $P[\Delta_1]$ and involve the same sequence of steps. Thus,

$$P[\Delta_3] = P[E_2] - \{P[E_1E_2] + P[E_2E_3] - P[E_1E_2E_3]\},$$
$$P[\Delta_7] = P[E_3] - \{P[E_1E_3] + P[E_2E_3] - P[E_1E_2E_3]\},$$

The computations of $P[\Delta_2]$, $P[\Delta_5]$, and $P[\Delta_6]$ are also quite similar and involve applying Equation 1.5-5. Thus,

$$P[\Delta_2] = P[E_1E_2] - P[E_1E_2E_3],$$
$$P[\Delta_5] = P[E_1E_3] - P[E_1E_2E_3],$$
$$P[\Delta_6] = P[E_2E_3] - P[E_1E_2E_3],$$

and finally,

$$P[\Delta_4] = P[E_1E_2E_3].$$

Now, recalling that $P[\bigcup_{i=1}^{i=3} E_i] = \sum_{i=1}^{7} P[\Delta_i]$, we merely add all the $P[\Delta_i]$ to obtain the desired result. This gives

$$P\left[\bigcup_{i=1}^{3} E_i\right] = \sum_{i=1}^{3} P[E_i] - \Big(P[E_1E_2] + P[E_1E_3] + P[E_2E_3]\Big) + P[E_1E_2E_3]. \qquad (1.5\text{-}10)$$

Note that this result makes sense because in adding the measures of the three events we have counted the double overlaps twice each. But if we subtract these three overlaps, we have not counted $E_1E_2E_3$ at all, and so must add it back in. If we adopt the notation $P_i \triangleq P[E_i]$, $P_{ij} \triangleq P[E_iE_j]$, and $P_{ijk} \triangleq P[E_iE_jE_k]$, where $1 \leq i < j < k \leq 3$, we can rewrite Equation 1.5-10 as

$$P\left[\bigcup_{i=1}^{3} E_i\right] = \sum_{i=1}^{3} P_i - \sum_{1\leq i<j\leq 3} P_{ij} + \sum_{1\leq i<j<k\leq 3} P_{ijk}.$$

The last sum contains only one term, namely P_{123}. Denote now each sum by the symbol S_l, where the l denotes the number of subscripts associated with the terms in that sum. Then

$$P\left[\bigcup_{i=1}^{3} E_i\right] = S_1 - S_2 + S_3, \text{ where } S_1 \triangleq \sum_{i=1}^{3} P_i,\ S_2 \triangleq \sum_{1\leq i<j\leq 3} P_{ij}, \text{ and}$$
$$S_3 \triangleq \sum_{1\leq i<j<k\leq 3} P_{ijk}.$$

Why this introduction of new notation? Using the symbols $S_l, l = 1, \ldots,$ we can extend Equation 1.5-10 to the general case.

Theorem 1.5-1 (*probability of union of n events*) The probability P that at least one among the events $E_1, E_2, \ldots, E_n$ occurs in a given experiment is given by

$$P = S_1 - S_2 + \ldots \pm S_n,$$

where $S_1 \triangleq \sum_{i=1}^{n} P_i$, $S_2 \triangleq \sum_{1\leq i<j\leq n}^{n} P_{ij}, \ldots,$ $S_n \triangleq \sum_{1\leq i<j<k<\ldots<l\leq n} P_{ijk\ldots l}$. The last sum has n subscripts and contains only one term. ■

The proof of this theorem is given in [1-8, p. 89]. It can also be proved by induction; that is, assume that $P = S_1 - S_2 + \ldots \pm S_n$ is true. Then show that for the case $n+1$, $P = S_1 - S_2 + \ldots \mp S_{n+1}$. We leave this exercise for the braver reader.

1.6 JOINT, CONDITIONAL, AND TOTAL PROBABILITIES; INDEPENDENCE

Assume that we perform the following experiment: We are in a certain U.S. city and wish to collect weather data about it. In particular we are interested in three events, call them A, B, and C, where

A is the event that on any particular day, the temperature equals or exceeds 10°C;
B is the event that on any particular day, the amount of precipitation equals or exceeds 5 millimeters;
C is the event that on any particular day A and B both occur, that is, $C \triangleq AB$.

Since C is an event, we can compute $P[C] = P[AB]$ and we call $P[AB]$ the *joint probability of the events A and B*. This notion can obviously be extended to more than two events; that is, $P[EFG]$ is the joint probability of events E, F, and G.[†] Now let n_E denote the number of days on which event E occurred. Over a thousand-day period ($n = 1000$), the following observations are made: $n_A = 811$, $n_B = 306$, $n_{AB} = 283$. By the relative frequency interpretation of probability

$$P[A] \simeq \frac{n_A}{n} = \frac{811}{1000} = 0.811,$$

$$P[B] \simeq \frac{n_B}{n} = 0.306,$$

$$P[AB] \simeq \frac{n_{AB}}{n} = 0.283.$$

Consider now the ratio n_{AB}/n_A. This would be the relative frequency with which event AB occurs given that event A occurs. Put into words, it is the fraction of time that the amount of precipitation equals or exceeds 5 millimeters on those days *given that* the temperature equals or exceeds 10°C. Thus, we are dealing with the frequency of an event given that or *conditioned upon the fact that another event has occurred.* Note that

$$\frac{n_{AB}}{n_A} = \frac{n_{AB}/n}{n_A/n} \simeq \frac{P[AB]}{P[A]}. \tag{1.6-1}$$

This empirical concept suggests that we introduce in our theory a *conditional* probability measure.

[†] E, F, G are any three events defined on the same probability space.

Conditional probability. The *conditional probability* $P[B|A]$ is defined by

$$P[B|A] \triangleq \frac{P[AB]}{P[A]}, \qquad \text{if } P[A] > 0, \tag{1.6-2}$$

and is read as "the probability that event B occurs given that event A has occurred." Similarly we have

$$P[A|B] \triangleq \frac{P[AB]}{P[B]}, \qquad \text{if } P[B] > 0. \tag{1.6-3}$$

Definitions 1.6-2 and 1.6-3 can be used to compute the joint probability of AB since

$$\begin{aligned} P[AB] &= P[A|B]P[B] \\ &= P[B|A]P[A]. \end{aligned}$$

Independence.

Definitions (*independence of events*) (i) Two events $A \in \mathscr{F}$, $B \in \mathscr{F}$ with $P[A] > 0$, $P[B] > 0$ are said to be independent *if and only if* (iff)

$$P[AB] = P[A]P[B]. \tag{1.6-4}$$

Since, in general, $P[AB] = P[B|A]P[A] = P[A|B]P[B]$ it follows that for independent events

$$P[A|B] = P[A], \tag{1.6-5a}$$

$$P[B|A] = P[B]. \tag{1.6-5b}$$

Thus, the definition satisfies our intuition: If A and B are independent, the outcome B should have no effect on the conditional probability of A and vice versa.
(ii) Three events A, B, C defined on $\mathscr{P}$ and having nonzero probabilities are said to be *jointly independent* iff

$$P[ABC] = P[A]P[B]P[C], \tag{1.6-6a}$$

$$P[AB] = P[A]P[B], \tag{1.6-6b}$$

$$P[AC] = P[A]P[C], \tag{1.6-6c}$$

$$P[BC] = P[B]P[C]. \tag{1.6-6d}$$

This is an extension of (i) above and suggests the pattern for the definition of n independent events $A_1, \ldots, A_n$. Note that it is 'not sufficient' to have just $P[ABC] = P[A]P[B]P[C]$. Pairwise independence must also be shown.
(iii) Let A_i, $i = 1, \ldots, n$, be n events contained in $\mathscr{F}$. The $\{A_i\}$ are said to be jointly independent iff

$$P[A_iA_j] = P[A_i]P[A_j]$$

$$P[A_iA_jA_k] = P[A_i]P[A_j]P[A_k]$$

$$\vdots$$

$$P[A_1 \dots A_n] = P[A_1]P[A_2]\dots P[A_n]$$

for all combination of indices such that $1 \le i < j < k < \dots \le n$. ■

Example 1.6-1

(*Sic bo*) The game *Sic bo* is played in gambling casinos. Players bet on the outcome of a throw of three dice. Many bets are possible each with a different payoff. We list two of them below with the associated payoffs in parentheses:

(1) *Specified* three of a kind (180 to 1), that is, pre-specified by the bettor;
(2) *Unspecified* three of a kind (30 to 1), that is, any three-way match.

What are the associated probabilities of winning from the bettor's point of view and his expected gain.

Solution

(1) (*specified three of a kind*) Let E_i be the event that the specified outcome appears on the ith toss. Then the event that three of a kind appear is $E_1E_2E_3$ with probability $P[E_1E_2E_3] = P[E_1]P[[E_2]P[E_3] = 1/216$, where we have used the fact that the three events are independent since they refer to different tosses. A fair payout would thus be 216 to 1, not 180 to 1.
(2) (*unspecified three of a kind*) On the first throw any number can come up. On the next two throws, numbers that match the first throw must come up. Hence P[three *unspecified*] $= 1 \times 1/6 \times 1/6 = 1/36$. A fair payout is thus 36 to 1, not 30 to 1.

Example 1.6-2

(*testing three events for independence*) An urn contains 10 numbered black balls (some even, some odd) and 20 numbered white balls (some even, some odd). Some of the balls of each color are lighter in weight than the others. The exact composition of the urn is shown in the tree diagram of Figure 1.6-1. The outcomes are triples ζ =(*color, weight, number*). The sample space Ω is the collection of all these triples. Each draw is completely random.

Let A denote the event of picking a black ball, B denote the event of picking a light ball, and C denote the event of picking an even-numbered ball. Are A, B, C independent events?

Solution We first test whether $P[ABC] = P[A]P[B]P[C]$. Now $P[A] = 1/3$ since 1/3 of the balls are black, $P[B] = 1/2$ since from the tree diagram we see that 15/30ths of the balls are light, and $P[C] = 2/5$ since 12/30 balls are even numbered. Now $P[ABC] = 2/30$ since the event ABC is black, light, and even and there are only two of them. Multiplying out we find that $P[ABC] = P[A]P[B]P[C]$. So the three events pass this part of the test

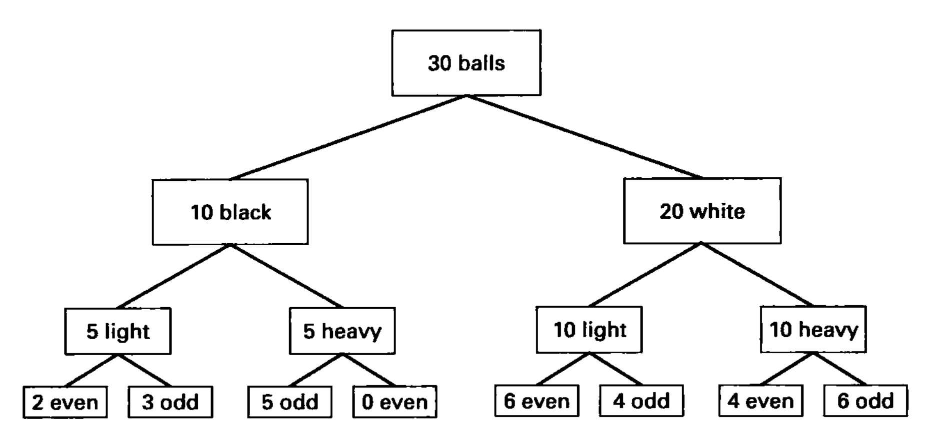


Figure 1.6-1 Diagram of composition of urn.

for independence. However, for (full) independence, we must also have $P[AB] = P[A]P[B]$, $P[AC] = P[A]P[C]$, and $P[BC] = P[B]P[C]$. Note that $P[AC] = 2/30$ while $P[A]P[C] = 1/3 \times 12/30 = 2/15 \neq 2/30$. Hence A, B, and C are not jointly independent.

Compound Experiments

Often we need to consider compound experiments or repeated trials. If we have a probability space defined for the individual experiments, we would like to see what this implies for the complete or compound experiment. There are two cases to consider, to model the physical fact that often the repeated trials seem to be independent of one another, while in other important cases the outcome seems to depend on the prior outcomes of earlier trials.

Independent experiments. Consider two independent experiments, meaning that the outcome of one is not affected by past, present, or future outcomes of the other. Let each have its own sample space Ω, outcomes ζ, events E, and probability measure P. Specifically, we have

$$\zeta_1 \in E_1 \subset \Omega_1 \text{ with measure } P_1 \quad \text{and} \quad \zeta_2 \in E_2 \subset \Omega_2 \text{ with measure } P_2,$$

as illustrated in Figure 1.6-2.

We want to be able to work with compound experiments, meaning that the sample space of the compound experiment is the *Cartesian product* of the two sample spaces,

$$\Omega \triangleq \Omega_1 \times \Omega_2$$

with vector outcomes (elements) $\boldsymbol{\zeta} = (\zeta_1, \zeta_2) \in E \subset \Omega$.

Example 1.6-3

(*flip two coins*) Let two experiments each consist of flipping a two-sided coin, with the two sides denoted H and T. Then we have $\Omega_1 = \{\text{H}, \text{T}\} = \Omega_2$. In the compound experiment, we have $\Omega = \{(\text{H}, \text{H}), (\text{H}, \text{T}), (\text{T}, \text{H}), (\text{T}, \text{T})\}$. We could also just as well write the outcomes $\boldsymbol{\zeta} \in \Omega$ as strings of characters H and T rather than vectors. In that notation, we have

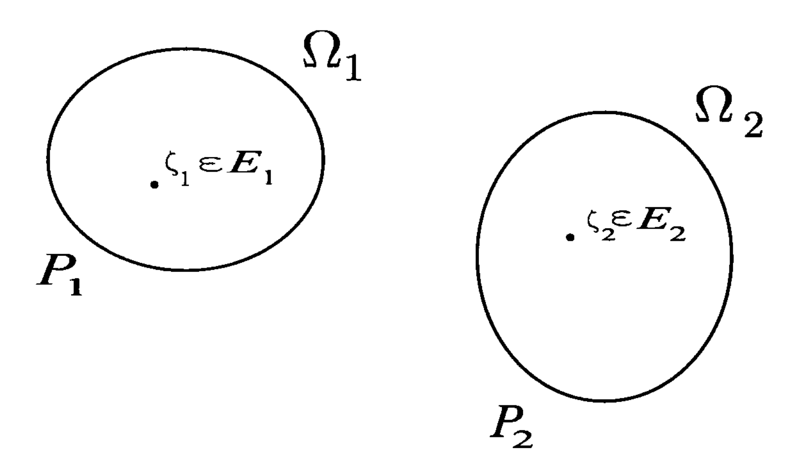


Figure 1.6-2 Two compound probabilistic experiments.

$\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. Considering event $E_1 = \{\text{T}\}$ in the first experiment, and event $E_2 = \{\text{H}\}$ in the second experiment, we have the event $E = \{\text{TH}\} = E_1 \times E_2 \subset \Omega$ in the compound experiment.

When we write a set with cross-product notation, we mean

$$E_1 \times E_2 \triangleq \{\boldsymbol{\zeta} = (\zeta_1, \zeta_1) | \zeta_1 \in E_1 \text{ and } \zeta_2 \in E_2\}.$$

So the elements in the cross-product of two sets are all the possible ordered pairs of elements, one from each set.

Example 1.6-4

(*toss two dice*) Let the two experiments now each consist of tossing a die, with the six faces (up) being denoted as outcomes 1–6. Then we have $\Omega_1 = \{1, 2, 3, 4, 5, 6\} = \Omega_2$. In the compound experiment, we have as outcomes the pair (or vector) elements of the cross-product sample space $\Omega = \{11, 12, ..., 16, 21, ..., 26, ..., 61, ..., 66\} = \Omega_1 \times \Omega_2$. Note that now all events (subsets of Ω) are not of the form $E_1 \times E_2$. In fact this is a special case. Consider the event $\{11, 12, 31\}$, for example. It is missing the outcome 32 contained in $\{1, 3\} \times \{1, 2\}$. However, we can write this event as a disjoint union over set cross products

$$\{11, 12, 31\} = \Big\{\{1\} \times \{1, 2\}\Big\} \cup \Big\{\{3\} \times \{1\}\Big\}.$$

Often we are interested in joint models for physical experiments that are independent of each other. This requires a definition. Thus, we define mathematically that two compound experiments are *independent* if the probabilities of events E can be expressed in terms of the individual probability measures P_1 and P_2.

Definition 1.6-1 Two experiments are said to be independent if (i) for a cross-product event $E = E_1 \times E_2$, we can write

$$P[E] \triangleq P_1[E_1]P_2[E_2],$$

(ii) the probability of a general event E in the compound experiment, can be written, in terms of singleton events, as

$$P[E] \triangleq \sum_{(\zeta_1,\zeta_2)\in E} P_1[\{\zeta_1\}]P_2[\{\zeta_2\}].$$

We can generalize this concept to combining n experiments to get the compound experiment's sample space

$$\Omega \triangleq \bigotimes_{i=1}^{n} \Omega_i$$
$$= \Omega_1 \times \Omega_2 \times \Omega_3 \times ... \times \Omega_n,$$

and vector (string) outcomes $\boldsymbol{\zeta} = (\zeta_1, ..., \zeta_n) \in E \subset \Omega$, the compound experiment's sample space. ■

Example 1.6-5

(*three experiments*) Consider three independent experiments, each with its own sample space $\Omega_i, i = 1, 2, 3..$ Let E_i be any arbitrary event in Ω_i. Then the general cross-product events $E = E_1 \times E_2 \times E_3$ in the compound experiment would have probabilities

$$P[E_1 \times E_2 \times E_3] = P_1[E_1]P_2[E_2]P_3[E_3],$$

where, the events E_i would be made up from unions and intersections of the measurable subsets of Ω_i.

Example 1.6-6

(*repeated coin flips*) Consider flipping a coin n times. Each flip can be considered a random, independent, experiment. Let the individual outcomes in each experiment be denoted H and T then the outcomes in the compound experiment are strings of H and T of length n. There are 2^n distinguishable ordered strings. The probability of a string having k H and $n-k$ T is given by

$$P[(\zeta_1, ..., \zeta_n)] = \prod_{i=1}^{n} P_i[\{\zeta_i\}]$$
$$= p^k q^{n-k},$$

where p and $q \triangleq 1 - p$, with $0 \leq p \leq 1$, are the individual probabilities of H and T, respectively, on a single coin flip.

We can also express these compound probabilities in terms of general events rather than singleton events. Again consider two experiments with probability spaces

$$\zeta_1 \in E_1 \subset \Omega_1 \text{ with measure } P_1 \quad \text{and} \quad \zeta_2 \in E_2 \subset \Omega_2 \text{ with measure } P_2;$$

then the compound experiment consists of the probability space

$$\zeta \triangleq (\zeta_1, \zeta_2) \in E \subset \Omega \text{ with measure } P,$$

where the compound probability measure P is defined for event $E \subset \Omega$ as follows. First, we must write the compound event in E as a disjoint union of cross-product events from the two experiments

$$E = \bigcup_{i=1}^{k} E_{1,i} \times E_{2,i},$$

for some positive integer k, where $E_{1,i}$ and $E_{2,i}$ are events in Ω_1 and Ω_2, respectively. In the simplest case E will itself be a cross-product event, and we will have $k = 1$, but as we have seen in Example 1.6-4, it will generally be necessary to take the union of several cross-product events to express an arbitrary event E in the compound sample space.

Definition 1.6-2 (alternative) Then when we say that the *experiments are independent*, we mean that for any event E in the compound experiment,

$$P[E] \triangleq \sum_{i=1}^{k} P_1[E_{1,i}]P_2[E_{2,i}], \quad \text{where} \quad E = \bigcup_{i=1}^{k} E_{1,i} \times E_{2,i},$$

a disjoint union, and where the $E_{1,i}$ and $E_{2,i}$ are events in Ω_1 and Ω_2, respectively. Here k is the number of cross-product events necessary to express compound event E.

We note that additivity of probability is appropriate since the events are disjoint. We can see immediately that this alternative definition is consistent with the definition in terms of elementary or singleton events given above. To see this simply take $E_{1,i}$ and $E_{2,i}$ as singleton events. Clearly this more general approach can also be extended to $n > 2$ experiments straightforwardly. We next turn to the more complicated case of multiple dependent experiments. ■

Dependent experiments.* Consider two "dependent experiments," meaning that the second experiment's probabilities will depend on the event that occurs in the first experiment. Let's say the first experiment consists of outcomes $\zeta_{1,i}$, where $i = 1, ..., k$, whose probabilities $P_1[\{\zeta_{1,i}\}]$ are given. The probability measures for the second experiment must be parametrized with index i from the first experiment, that is,

$$P_{2,i}[E_2] \text{ for each event } E_2 \subset \Omega_2,$$

where Ω_2 is the sample space for the second experiment. This is illustrated in Figure 1.6-3.

Then we write the probability measure for the compound experiment as follows.

Definition 1.6-3 (*dependent experiments*) Let $E_1 = \{\zeta_{1,i}\}$ be a singleton event in Ω_1 for some i, and let E_2 be an event in Ω_2; then consider the cross-product event $E = E_1 \times E_2$ in the compound experiment. We then write

$$P[E] \triangleq P_1[\{\zeta_{1,i}\}]P_{2,i}[E_2],$$

where the probability measure in the second experiment is a function of the outcome in the first experiment. ■

*Starred material can be omitted on a first reading.

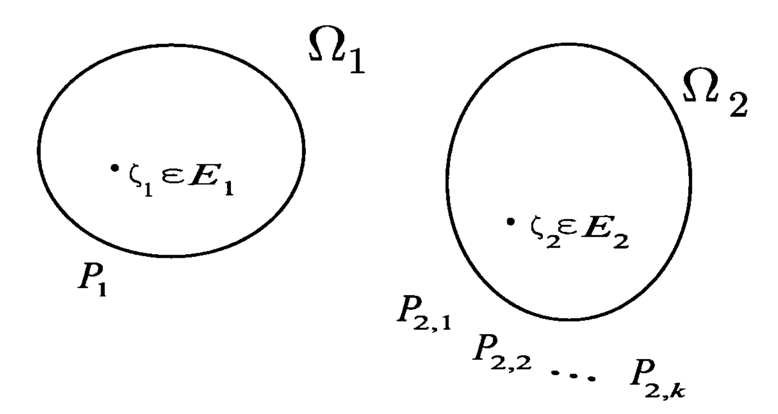


Figure 1.6-3 Two dependent compound experiments.

First we note that this definition is consistent with the definition above for the case of independent experiments. This is because in the case of independent events all the $P_{2,i}$ are the same, that is, $P_{2,i} = P_2$, for all i.

More generally, let the event in the first experiment be $E_1 = \bigcup_i \{\zeta_{1,i}\}$, that is, the union of i elementary (singleton) events; then the probability of the compound event $E = E_1 \times E_2$ is written as

$$P[E] \triangleq \sum_i P_1[\{\zeta_{1,i}\}]P_{2,i}[E_2].$$

Here additivity makes sense since only one of the i elementary events $\{\zeta_{1,i}\}$ can occur in the first experiment.

Example 1.6-7

(*flip biased coins*) Let there be three biased coins considered. We flip the first one, with $p_1 = P_1[\{\text{H}\}]$. Depending on the outcome, H or T, we then flip coin 2 or coin 3, respectively. Assume for coin 2 that the probability $p_2 = P_2[\{\text{H}\}]$, and that for coin 3, we have $p_3 = P_3[\{\text{H}\}]$. Here, of course, we assume that all the p_i satisfy $0 < p_i < 1$. Then for $p_2 \neq p_3$, we have the case of dependent experiments. Computing, for example, $P\{\text{HT}\}$, we get $p_1(1-p_2)$ and for $P\{\text{TH}\}$, we get $(1-p_1)p_3$, etc.

Example 1.6-8

(*conditioning on events*) Consider that the weather today can be sunny, cloudy, or rainy with probabilities $p_{1,s}, p_{1,c}$, and $p_{1,r}$, respectively, where these three sum to one. Then tomorrow, it may be also sunny, cloudy, or rainy, and that may depend on what happened today. So the conditional probability for the weather tomorrow can depend on these conditioning events, and would be expected to be a different measure for each one. We would have a set of three conditional probability measures for day 2, one for each condition from day 1.

Relation to conditional probability. Consider a compound experiment with two component experiments $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ that are *independent*, so that we have the compound experiment $(\Omega, \mathcal{F}, P)$ with

$$P[E_1 \times E_2] = P_1[E_1]P_2[E_2],$$

for all cross-product events $E_1 \times E_2 \in \mathcal{F}$, where $E_1 \in \mathcal{F}_1$ and $E_2 \in \mathcal{F}_2$. We can think of the first experiment as occurring before the second one. Let the conditioning event $B \in \mathcal{F}$ be of the form $B = B_1 \times \Omega_2$, where $B_1 \in \mathcal{F}_1$. Then $P[B] = P_1[B_1] \cdot 1$. Similarly, let the event $A \in \mathcal{F}$ be of the form $A = \Omega_1 \times A_2$, where $A_2 \in \mathcal{F}_2$; then $P[A] = 1 \cdot P_2[A_2]$, and we find that the conditional probability

$$\begin{aligned} P[A|B] &= \frac{P[AB]}{P[B]} \\ &= \frac{P[(\Omega_1 \times A_2) \cap (B_1 \times \Omega_2)]}{P_1[B_1]} \\ &= \frac{P[B_1 \times A_2]}{P_1[B_1]} \\ &= \frac{P_1[B_1]P_2[A_2]}{P_1[B_1]} \\ &= P_2[A_2], \end{aligned}$$

where we have noted that

$$\begin{aligned} (\Omega_1 \times A_2) \cap (B_1 \times \Omega_2) &= \{(\zeta_1, \zeta_2)|\zeta_2 \in A_2\} \cap \{(\zeta_1, \zeta_2)|\zeta_1 \in B_1\} \\ &= B_1 \times A_2. \end{aligned}$$

Now this is what we expect to happen for two independent experiments. But, what happens when the two experiments are dependent?

***Example 1.6-9**

(*dependent case*) Consider a compound experiment with two components as above, that is, $B = B_1 \times \Omega_2$ and $A = \Omega_1 \times A_2$, but now assume that these experiments are dependent. Assume the number of outcomes in the first experiment to be a finite number k and write the probability measure of the second experiment as a function of the outcome on the first experiment, that is, $P_{2,i}$ for each outcome $\zeta_{1,i} \in \Omega_1$ for $i = 1, ..., k$. Assume also that $B_1 = \{\zeta_{1,i}\}$ for some value i. Then proceeding as in the last example, we have

$$\begin{aligned} P[A|B] &= \frac{P[(\Omega_1 \times A_2) \cap (B_1 \times \Omega_2)]}{P_1[B_1]} \\ &= \frac{P_1[B_1]P_{2,i}[A_2]}{P_1[B_1]} \\ &= P_{2,i}[A_2], \quad \text{as expected.} \end{aligned}$$

Example 1.6-10

(*communication channel and source*) In a binary communication system, we have a binary source S along with a binary channel C (Figure 1.6-4) defined in terms of its conditional probabilities. The sample space Ω for this combined experiment is $\Omega = \{\zeta = (x, y) \colon x =$ and $y = 0$ or $1\} = \{(0,0), (0,1), (1,0), (1,1)\}$, where x denotes the source output that is the channel input, and y denotes the channel output. The joint probability function is then given as $P[\{(x,y)\}] = P_S[\{x\}]P_C[\{y\}|\{x\}]\ x, y = 0, 1$, where P_S is the probability measure of the source S and P_C is the conditional probability measure of the channel C.

Because of noise a transmitted zero sometimes gets decoded as a received one and vice versa. From repeated use of the channel, it is known that

$$P_C[\{0\}|\{0\}] = 0.9, \qquad P_C[\{1\}|\{0\}] = 0.1,$$

$$P_C[\{0\}|\{1\}] = 0.1, \qquad P_C[\{1\}|\{1\}] = 0.9,$$

and by design of the source $P_S[\{0\}] = P_S[\{1\}] = 0.5$.† The various probabilities of the singleton events in the joint experiment are then

$$P[\{(0,0)\}] = P_C[\{0\}|\{0\}]P_S[\{0\}] = 0.45$$

$$P[\{(0,1)\}] = P_C[\{1\}|\{0\}]P_S[\{0\}] = 0.05$$

$$P[\{(1,0)\}] = P_C[\{0\}|\{1\}]P_S[\{1\}] = 0.05$$

$$P[\{(1,1)\}] = P_C[\{1\}|\{1\}]P_S[\{1\}] = 0.45.$$

We can also define some events on the compound or combined sample space

$$X_0 \triangleq \text{"event that } x = 0\text{"} \quad \text{and } X_1 \triangleq \text{"event that } x = 1\text{"}$$

$$Y_0 \triangleq \text{"event that } y = 0\text{"} \quad \text{and } Y_1 \triangleq \text{"event that } y = 1\text{"}$$

and rewrite the above channel conditional probabilities as

$$P[Y_0|X_0] = 0.9 \quad \text{and} \quad P[Y_1|X_0] = 0.1$$

$$P[Y_0|X_1] = 0.1 \quad \text{and} \quad P[Y_1|X_1] = 0.9.$$

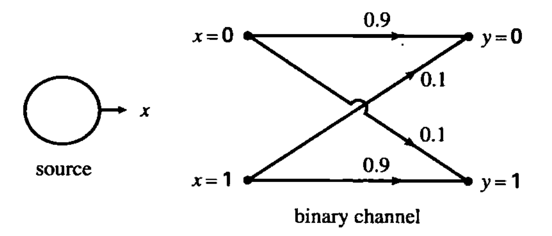


Figure 1.6-4 A binary communication system.

†It is good practice to design a code in which the zeros and ones appear at close to the same rate since this puts the signaling capacity of the channel to greatest use.

The source probabilities are then just expressed as

$$P[X_0] = 0.5 \quad \text{and} \quad P[X_1] = 0.5.$$

In the combined experiment, the above joint probabilities become

$$P[X_0 \cup Y_0] = P[Y_0|X_0]P[X_0] = 0.45$$
$$P[X_0 \cup Y_1] = P[Y_1|X_0]P[X_0] = 0.05$$
$$P[X_1 \cup Y_0] = P[Y_0|X_1]P[X_1] = 0.05$$
$$P[X_1 \cup Y_1] = P[Y_1|X_1]P[X_0] = 0.45.$$

The introduction of conditional probabilities raises the important question of whether conditional probabilities satisfy Axioms 1 to 3. In other words, given any two events E, F such that $EF = \phi$ and a third arbitrary event A with $P[A] > 0$, all belonging to the σ-field of events $\mathscr{F}$ in the probability space $(\Omega, \mathscr{F}, P)$, does

$$P[E|A] \geq 0?$$
$$P[\Omega|A] = 1?$$
$$P[E \cup F|A] = P[E|A] + P[F|A] \quad \text{for } EF = \phi?$$

The answer is yes. We leave the details as an exercise to the reader. They follow directly from the definition of conditional probability and the three Kolmogorov axioms.

Example 1.6-11

(*probability trees*) Three events A, B, and C are often specified in terms of conditional probabilities as follows: ·

$$P[A], P[B|A], P[B|A^c] \text{ and}$$
$$P[C|BA], P[C|BA^c], P[C|B^cA], P[C|B^cA^c].$$

In such a case the problem can be summarized in a tree diagram, such as Figure 1.6-5, where the branches are labeled with the relevant conditional probabilities and the node values are the corresponding joint probabilities. Here the root node can be thought of as having value 1.0 and being associated with the certain event Ω. If we want to evaluate the probability of an event on a leaf (the last set of nodes) of the tree, we just multiply the conditional probabilities on its path.

A way this can arise is if the events come from compound experiments conducted sequentially, so that the event B depends on the event A, and in turn the event C depends on them both. A more general tree would have more than two outgoing branches at each node indicating more than two events were possible, for example, $A_1, A_2, ..., A_N$. The conditional probabilities can be stored in a data structure in a machine, which could be queried for answers to various joint probability questions, such as: What is the probability of the joint

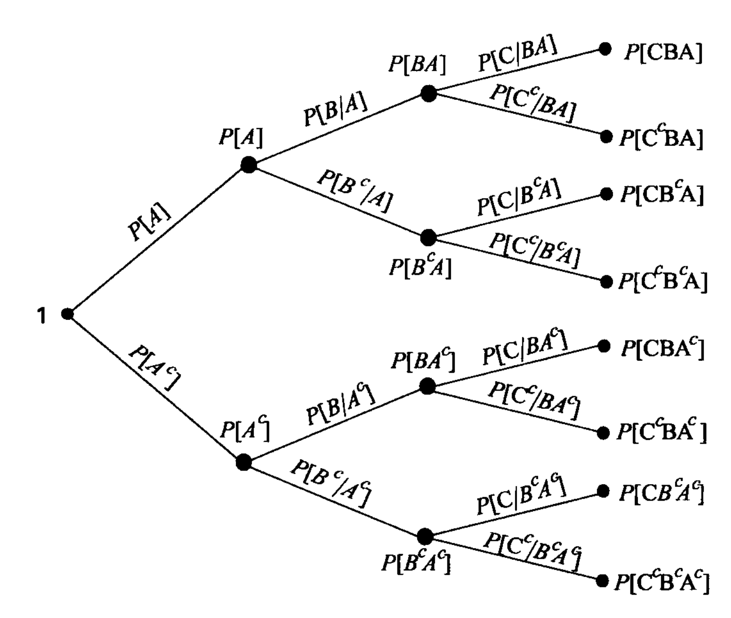


Figure 1.6-5 A probability tree diagram with conditional probabilities on the branches and joint probabilities at the nodes.

event $C_k B_l A_n$ which could be answered by tracing the corresponding path in the stored data structure and then multiplying the values on its branches? For a concrete example, take first round A_n events to indicate the health (good, fair, poor) of a plant purchased at a local nursery, then B_l can indicate its health one week later, and C_k can indicate the health at two weeks from purchase.

The next example, illustrating the use of joint and conditional probabilities, has applications in real life where we might be forced to make important decisions without knowing all the facts.

Example 1.6-12

(*beauty contest*)† Assume that a beauty contest is being judged by the following rules: (1) There are N contestants not seen by the judges before the contest, and (2) the contestants are individually presented to the judges in a random sequence. Only one contestant appears before the judges at any one time. (3) The judges must decide on the spot whether the contestant appearing before them is the most beautiful. If they decide in the affirmative, the contest is over but the risk is that a still more beautiful contestant is in the group as yet not displayed. In that case the judges would have made the wrong decision. On the other hand, if they pass over the candidate, the contestant is disqualified from further consideration even if it turns out that all subsequent contestants are less beautiful. What is a good

†Thanks are due to Geof Williamson and Jerry Tiemann for valuable discussions regarding this problem.

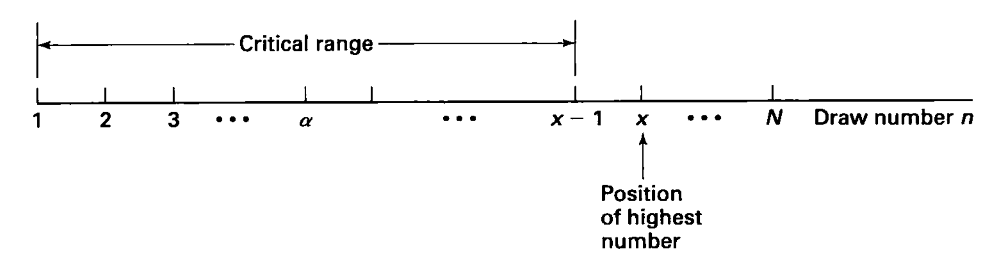


Figure 1.6-6 The numbers along the axis represent the chronology of the draw, *not* the number actually drawn from the bag.

strategy to follow to increase the probability of picking the most beautiful contestant over that of a random choice?

Solution To make the problem somewhat more quantitative, assume that all the virtues of each contestant are summarized into a single "beauty" number. Thus, the most beautiful contestant is associated with the highest number, and the least beautiful has the lowest number. We make no assumptions regarding the distribution or chronology of appearance of the numbers. The numbers, unseen by the judges, are placed in a bag and the numbers are drawn individually from the bag. We model the problem then as one of randomly drawing the "beauty" numbers from a bag. We consider that the draws are ordered along a line as shown in Figure 1.6-6. Thus, the first draw is number 1, the second is 2, and so forth. At each draw, a number appears. Is it the largest of all the N numbers?

Assume that the following "wait-and-see" strategy is adopted: We pass over the first k draws (i.e., we reject the first k contestants) but record the highest number (i.e., the most beautiful contestant) observed within this group of k. Then we continue drawing numbers (i.e., call for more contestants to appear). The first draw (contestant) after the k passed-over draws that yields a number exceeding the largest number from the first k draws is taken to be the winner. If a larger number does not occur, then the judge declines to vote and we count this as an error.

Let us define $E_j(k)$ as the event that the *largest number* that is drawn from the first j draws occurs in the group of first k draws. Then for $j \leq k$, $E_j(k) = \Omega$ (the certain event), but for $j > k$, $E_j(k)$ will be a proper subset of Ω. Let x denote the draw that will contain the largest number among the N numbers in the bag. Then two events must occur jointly for the correct decision to be realized. (1) (obvious) $\{x > k\}$; and (2) (subtle) $E_j(k)$ for all j such that $k < j < x$. Then for a correct decision C to happen, the subevent $\{x = j + 1\}$ must occur jointly with the event $E_j(k)$ for each j such that $k < j < N$. The event $\{x > k\}$ can be resolved into disjoint subevents as

$$\{x > k\} = \{x = k + 1\} \cup \{x = k + 2\} \cup \ldots \cup \{x = N\}.$$

Thus,

$$C = \{x = k + 1, E_k(k)\} \cup \{x = k + 2, E_{k+1}(k)\} \ldots \cup \{x = N, E_{N-1}(k)\},$$

and the probability of a correct decision is

$$P[C] = \sum_{j=k}^{N-1} P[x = j+1, E_j(k)], \qquad \text{because these events are disjoint,}$$

$$= \sum_{j=k}^{N-1} P[E_j(k)|x = j+1]P[x = j+1]$$

$$= \frac{1}{N}\sum_{j=k}^{N-1}\frac{k}{j},$$

where we have used the fact that $P[x = j+1] = \frac{1}{N}$ since all N draws are equally likely to result in the largest number. Also $P[E_j(k)|x = j+1] = \frac{k}{j}$ since the "largest" draw from the first j draws could equally likely be any of the first j draws, and so the probability that it is in the first k of these j draws is given by the fraction $\frac{k}{j}$.

By the Euler summation formula† for large N

$$P[C] = \frac{1}{N}\sum_{j=k}^{N-1}\frac{k}{j}$$

$$= \frac{k}{N}\left(\frac{1}{k} + \frac{1}{k+1} + \frac{1}{k+2} + \dots + \frac{1}{N-1}\right)$$

$$\simeq \frac{k}{N}\int_k^N \frac{dx}{x} \quad \text{for} \quad k \text{ large enough,}$$

$$= \frac{k}{N}\ln\frac{N}{k}.$$

Neglecting the integer constraint, an approximate best choice of k, say k_0, can be found by differentiation. Setting

$$\frac{dP[C]}{dk} = 0,$$

we find that

$$k_0 \simeq \frac{N}{e}.$$

Invoking the integer constraint we round k_0 to the nearest integer, as to finally obtain

$$k_0 \simeq \left\lfloor \frac{N}{e} + \frac{1}{2} \right\rfloor,$$

†See, for example, G. F. Carrier et al., *Functions of a Complex Variable* (New York: McGraw-Hill, 1966), p. 246, or visit the Wikipedia page: *Euler–Maclaurin formula* (http://en.wikipedia.org/wiki/Euler%E2%80%93Maclaurin_formula).

where $\lfloor \cdot \rfloor$ denotes the least-integer function. The maximum probability of a correct decision $P[C]$ then becomes

$$\begin{aligned} P[C] &\simeq \frac{\lfloor \frac{N}{e} + \frac{1}{2} \rfloor}{N} \ln \frac{N}{\lfloor \frac{N}{e} + \frac{1}{2} \rfloor} \\ &\simeq \frac{1}{e} \ln e \doteq 0.367. \end{aligned}$$

Thus, we should let approximately the first third (more precisely 36.7 percent) of the contestants pass by before beginning to judge the contestants in earnest. We assume that N is reasonably large for this result to hold. The interesting fact is that the result is independent of (large) N while the probability of picking the most beautiful candidate by random selection decreases as $1/N$.

Here are some other situations that require a strategy that will maximize the probability of making the right decision.

1. You are apartment-hunting and have selected 30 rent-controlled flats to inspect. You see an apartment that you like but you are not ready to make an offer because you think that the next apartment to be shown might be more desirable. However, none of the subsequent apartments that you visit measure up to the first. Sadly, your offer for that apartment is rejected because, meanwhile, someone else rented it. You will have to settle for a far lesser desirable apartment because you hesitated.
2. You are looking for a partner to spend the rest of your life with. To that end you contract with a singles dating agency to meet 50 possible life partners at the rate of one date per week. On your ninth date, you decide that you have found your life's partner and offer marriage, which is accepted. However, you forget to tell the dating agency to stop introducing you to additional partners. The following week you are introduced to a date that in all qualities surpasses your chosen one. You kick yourself for having acted too impulsively.
3. You are interviewing candidates for a high-level position in the government. To reduce the possibility of discrimination on your part you are bound by the following rules: You are to interview the candidates in sequence and offer the job to the first candidate who is qualified according to the job description. If you reject a candidate it means that he/she was not qualified and so you must state in writing in your report. However, you are savvy enough to know that even among the qualified candidates there will be those that are superbly qualified while others will be merely qualified. You want to hire the best person for the job. What should your strategy be?

Total Probability. In many problems in engineering and science we would like to compute the unconditional probability $P[B]$ of an event B in terms of the sum of weighted conditional probabilities. Such a computation is easily realized through the following theorem.

Theorem 1.6-1 Let $A_1, A_2, \ldots, A_n$ be n mutually exclusive events such that $\bigcup_{i=1}^{n} A_i = \Omega$ (the A_i's are *exhaustive*). Let B be any event defined over the probability space of the A_i's. Then, with $P[A_i] \neq 0$ all i,

$$P[B] = P[B|A_1]P[A_1] + \ldots + P[B|A_n]P[A_n]. \tag{1.6-7}$$

Sometimes $P[B]$ is called the *total probability* of B because the expression on the right is a weighted average of the conditional probabilities of B.

Proof We have $A_iA_j = \phi$ for all $i \neq j$ and $\bigcup_{i=1}^{n} A_i = \Omega$. Also $B\Omega = B = B\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} BA_i$. But by definition of the intersection operation, $BA_i \subset A_i$; hence $(BA_i)(BA_j) = \phi$ for all $i \neq j$. Thus, from Axiom 3 (generalized to n events):

$$\begin{aligned} P[B] = P\left[\bigcup_{i=1}^{n} BA_i\right] &= P[BA_1] + P[BA_2] + \ldots + P[BA_n] \\ &= P[B|A_1]P[A_1] + \ldots + P[B|A_n]P[A_n]. \end{aligned} \tag{1.6-8}$$

The last line follows from Equation 1.6-2. ■

Example 1.6-13

(*more on binary channel*) For the binary communication system shown in Figure 1.6-4, compute the unconditional output probabilities $P[Y_0]$ and $P[Y_1]$.

Solution Continuing with the notation of binary communication Example 1.6-10, we use Equation 1.6-8 as follows:

$$\begin{aligned} P[Y_0] &= P[Y_0|X_0]P[X_0] + P[Y_0|X_1]P[X_1] \\ &= P_C[0|0]P_S[0] + P_C[0|1]P_S[1]^{\dagger} \\ &= (0.9)(0.5) + (0.1)(0.5) \\ &= 0.5. \end{aligned}$$

We can compute $P[Y_1]$ in a similar fashion or by noting that $Y_0 \cup Y_1 = \Omega$ and $Y_0 \cap Y_1 = \phi$; that is, they are disjoint. Hence $P[Y_0] + P[Y_1] = 1$, implying $P[Y_1] = 1 - P[Y_0] = 0.5$.

1.7 BAYES' THEOREM AND APPLICATIONS

The previous results enable us now to write a fairly simple formula known as Bayes' theorem.‡ Despite its simplicity, this formula is widely used in biometrics, epidemiology, and communication theory.

† For notational ease, we have abbreviated these terms by leaving off the curly brackets. We retain the square brackets for probabilities P through to remind that they are set functions.

‡Named after Thomas Bayes, English mathematician/philosopher (1702–1761).

Bayes' Theorem Let A_i, $i = 1, \ldots, n$, be a set of disjoint and exhaustive events defined on a probability space $\mathscr{P}$. Then, $\bigcup_{i=1}^{n} A_i = \Omega$, $A_i A_j = \phi$ for all $i \neq j$. With B any event defined on $\mathscr{P}$ with $P[B] > 0$ and $P[A_i] \neq 0$ for all i

$$P[A_j|B] = \frac{P[B|A_j]P[A_j]}{\sum_{i=1}^{n} P[B|A_i]P[A_i]}. \tag{1.7-1}$$

Proof The denominator is simply $P[B]$ by Equation 1.6-8 and the numerator is simply $P[A_jB]$. Thus, Bayes' theorem is merely an application of the definition of conditional probability. ■

Remark In practice the terms in Equation 1.7-1 are given various names: $P[A_j|B]$ is known as the *a posteriori* probability of A_j given B; $P[B|A_j]$ is called the *a priori* probability of B given A_j; and $P[A_i]$ is the *causal* or *a priori* probability of A_i. In general *a priori* probabilities are estimated from past measurements or presupposed by experience while *a posteriori* probabilities are measured or computed from observations.

Example 1.7-1

(*inverse binary channel*) In a communication system a zero or one is transmitted with $P_S[0] = p_0$, $P_S[1] = 1 - p_0 \triangleq p_1$, respectively. Due to noise in the channel, a zero can be received as a one with probability β, called the *cross-over probability*, and a one can be received as a zero also with probability β. A one is observed at the output of the channel. What is the probability that a one was output by the source and input to the channel, that is, transmitted?

Solution The structure of the channel is shown in Figure 1.7-1. We write

$$P[X_1|Y_1] = \frac{P[X_1Y_1]}{P[Y_1]} \tag{1.7-2}$$

$$= \frac{P_C[1|1]P_S[1]}{P_C[1|1]P_S[1] + P_C[1|0]P_S[0]} \tag{1.7-3}$$

$$= \frac{p_1(1-\beta)}{p_1(1-\beta) + p_0\beta}. \tag{1.7-4}$$

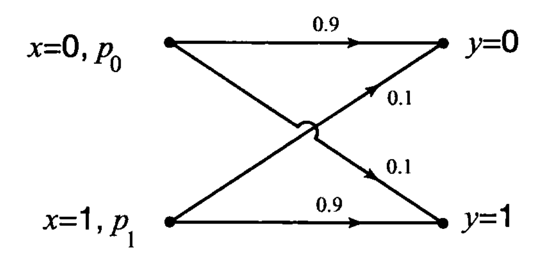


Figure 1.7-1 Representation of a binary communication channel subject to noise.

If $p_0 = p_1 = \frac{1}{2}$, the inverse or *a posteriori* probability $P[X_1|Y_1]$ depends on β as shown in Figure 1.7-2. The channel is said to be noiseless if $\beta = 0$, but notice that the channel is just as useful when $\beta = 1$. Just invert the outputs in this case!

Example 1.7-2

(*amyloid test for Alzheimer's disease*) On August 10, 2010 there was a story on network television news that a promising new test was developed for Alzheimer's disease. It was based on the occurrence of the protein amyloid in the spinal (and cerebral) fluid, which could be detected via a spinal tap. It was reported that among Alzheimer's patients (65 and older) there were 90 percent who had amyloid protein, while among the Alzheimer's free group (65 and older) amyloid was present in only 36 percent of this subpopulation. Now the general incidence of Alzheimer's among the group 65 and older is thought to be 10 percent from various surveys over the years. From this data, we want to find out: Is it really a good test?

First we construct the probability space for this experiment. We set $\Omega = \{00, 01, 10, 11\}$ with four outcomes:

00 = "no amyloid" and "no Alzheimer's,"

01 = "no amyloid" and "Alzheimer's,"

10 = "amyloid" and "no Alzheimer's."

11 = "amyloid" and "Alzheimer's."

On this sample space, we define two events: $A \triangleq \{10, 11\}$ = "amyloid" and $B \triangleq \{01, 11\}$= "Alzheimer's." From the data above we have

$$P[A|B] = 0.9 \quad \text{and} \quad P[A|B^c] = 0.36.$$

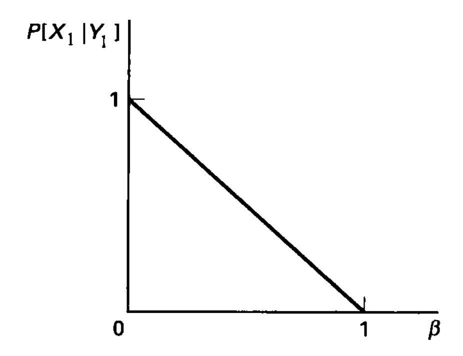


Figure 1.7-2 *A posteriori* probability versus β.

Also from the general population (65 and greater), we know

$$P[B] = 0.1 \quad \text{and} \quad P[B^c] = 1 - P[B] = 0.9.$$

Now to determine if the test is good, we must look at the situation after we give the test, and this is modeled by the conditional probabilities after the test. They are either $P[\cdot|A]$, if the test is positive for amyloid, or $P[\cdot|A^c]$, if the test is negative. So, we can use total probability to find answers such as $P[B|A]$. We have

$$\begin{aligned} P[B|A] &= \frac{P[A|B]P[B]}{P[A|B]P[B] + P[A|B^c]P[B^c]} \\ &= \frac{0.9 \times 0.1}{0.9 \times 0.1 + 0.36 \times 0.9} \\ &= 0.217. \end{aligned}$$

So, among the group that tests positive, only about 22 percent will actually have Alzheimer's. The test does not seem so promising now. Why is this? Well, the problem is that we are never in the "knowledge state" characterized by event B where conditional probability $P[\cdot|B]$ is relevant. Before the test is given, our knowledge state is characterized by the unconditional probabilistic knowledge $P[\cdot]$. After the test, we have knowledge state determined by whether event A or A^c has occurred; that is, our conditional probabilistic state is either $P[\cdot|A]$ or $P[\cdot|A^c]$. You see, we enter into states of knowledge either "given A" or "given A^c" by testing the population. So we are never in the situation or knowledge state where $P[\cdot|B]$ or $P[\cdot|B^c]$ is the relevant probability measure. So the given information $P[A|B] = 0.9$ and $P[A|B^c] = 0.36$ is not helpful to *directly* decide whether the test is useful or not. This is the logical fallacy of reasoning with $P[A|B]$ instead of $P[B|A]$, but there is another very practical thing going on here too in this particular example.

When we calculate $P[B^c|A] = 1.0 - 0.217 = 0.783$, this means that about 78 percent of those with positive amyloid tests do not have Alzheimer's. So the test is not useful due to its high false-positive rate. Again, as in the previous example, the scarcity of Alzheimer's in the general population (65 and greater) is a problem here, and any test will have to overcome this in order to become a useful test.

1.8 COMBINATORICS†

Before proceeding with our study of basic probability, we introduce a number of counting formulas important for counting equiprobable events. Some of the results presented here will have immediate application in Section 1.9; others will be useful later.

†This material closely follows that of William Feller [1-8].

A *population of size* n will be taken to mean a collection (set) of n elements without regard to order. Two populations are considered different if one contains at least one element not contained in the other. A *subpopulation* of size r from a population of size n is a subset of r elements taken from the original population. Likewise, two subpopulations are considered different if one has at least one element different from the other.

Consider a population of n elements $a_1, a_2, \ldots, a_n$. Any ordered arrangement $a_{k_1}, a_{k_2}, \ldots, a_{k_r}$ of r symbols is called an *ordered sample* of size r. Consider now the generic urn containing n distinguishable numbered balls. Balls are removed one by one. How many different ordered samples of size r can be formed? There are two cases:

(*i*) *Sampling with replacement.* Here after each ball is removed, its number is recorded and it is returned to the urn. Thus, for the first sample there are n choices, for the second there are again n choices, and so on. Thus, we are led to the following result: For a population of n elements, there are n^r different ordered samples of size r that can be formed with replacement.

(*ii*) *Sampling without replacement.* After each ball is removed, it is not available anymore for subsequent samples. Thus, n balls are available for the first sample, $n-1$ for the second, and so forth. Thus, we are now led to the result: For a population of n elements, there are

$$(n)_r \triangleq n(n-1)(n-2)\ldots(n-r+1)$$
$$= \frac{n!}{(n-r)!} \tag{1.8-1}$$

different ordered samples of size r that can be formed without replacement†

The Number of Subpopulations of Size r in a Population of Size n. A basic problem that often occurs in probability is the following: How many groups, that is, subpopulations, of size r can be formed from a population of size n? For example, consider six balls numbered 1 to 6. How many groups of size 2 can be formed? The following table shows that there are 15 groups of size 2 that can be formed:

12	23	34	45	56
13	24	35	46	
14	25	36		
15	26			
16				

Note that this is different from the number of ordered samples that can be formed without replacement. These are $(6 \cdot 5 = 30)$:

12	21	31	41	51	61
13	23	32	42	52	62
14	24	34	43	53	63
15	25	35	45	54	64
16	26	36	46	56	65

†Different samples will often contain the same subpopulation but with a different *ordering*. For this reason we sometimes speak of $(n)_r$ ordered samples that can be formed without replacement.

Also it is different from the number of samples that can be formed with replacement ($6^2 = 36$):

11	21	31	41	51	61
12	22	32	42	52	62
13	23	33	43	53	63
14	24	34	44	54	64
15	25	35	45	55	65
16	26	36	46	56	66

A general formula for the number of subpopulations, C_r^n of size r in a population of size n can be computed as follows: Consider an urn with n distinguishable balls. We already know that the number of ordered samples of size r that can be formed is $(n)_r$. Now consider a specific subpopulation of size r. For this subpopulation there are $r!$ arrangements and therefore $r!$ different ordered samples. Thus, for C_r^n subpopulations there must be $C_r^n \cdot r!$ different ordered samples of size r. Hence

$$C_r^n \cdot r! = (n)_r$$

or

$$C_r^n = \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!} \triangleq \binom{n}{r}. \tag{1.8-2}$$

Equation 1.8-2 is an important result, and we shall apply it in the next section. The symbol

$$C_r^n \triangleq \binom{n}{r}$$

is called a *binomial coefficient.* Clearly

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{(n-r)!r!} \triangleq \binom{n}{n-r} = C_{n-r}^n. \tag{1.8-3}$$

We already know from Section 1.4 that the total number of subsets of a set of size n is 2^n. The number of subsets of size r is $\binom{n}{r}$. Hence we obtain that

$$\sum_{r=0}^{n} \binom{n}{r} = 2^n.$$

A result which can be viewed as an extension of the binomial coefficient C_r^n is given by the following.

Theorem 1.8-1 Let $r_1, \ldots, r_l$ be a set l of nonnegative integers such that $r_1 + r_2 + \ldots + r_l = n$. Then the number of ways in which a population of n elements can be partitioned

into l subpopulations of which the first contains r_1 elements, the second r_2 elements, and so forth, is

$$\frac{n!}{r_1!r_2!\dots r_l!} \, . \tag{1.8-4}$$

This coefficient is called the *multinomial coefficient.* Note that the order of the subpopulation is essential in the sense that ($r_1 = 7$, $r_2 = 10$) and ($r_1 = 10$, $r_2 = 7$) represent different partitions. However, the order within each group does not receive attention. For example, suppose we have five distinguishable balls (1,2,3,4,5) and we ask how many subpopulations can be made with three balls in the first group and two in the second. Here $n = 5$, $r_1 = 3$, $r_2 = 2$, and $r_1 + r_2 = 5$. The answer is $5!/3!2! = 10$ and the partitions are

Group 1:	1,2,3	2,3,4	3,4,5	4,5,1	5,1,2	2,4,5	2,3,5	1,3,5	1,3,4	1,2,4
Group 2:	4,5	5,1	1,2	2,3	3,4	1,3	1,4	2,4	2,5	3,5

Note that the order is important in that had we set $r_1 = 2$ and $r_2 = 3$ we would have gotten a *different* partition, for example,

Group 1:	4,5	5,1	1,2	2,3	3,4	1,3	1,4	2,4	2,5	3,5
Group 2:	1,2,3	2,3,4	3,4,5	4,5,1	5,1,2	2,4,5	2,3,5	1,3,5	1,3,4	1,2,4

The partition (4,5), (1,2,3) is, however, identical with (5,4), (2,1,3).

Proof Note that we can rewrite Equation 1.8-4 as

$$\frac{n!}{r_1!}\cdot\frac{1}{r_2!}\cdot\dots\cdot\frac{1}{r_l!}\,.$$

$$= \frac{n!}{r_1!(n-r_1)!}\cdot\frac{(n-r_1)!}{r_2!(n-r_1-r_2)!}\,\frac{(n-r_1-r_2)!}{r_3!(n-r_1-r_2-r_3)!}\cdots\frac{\left(n-\sum_{j=1}^{l-1}r_j\right)!}{r_l!\left(n-\sum_{j=1}^{l}r_j\right)!}.$$

Recalling that $0! = 1$, we see that the last term is unity. Then the multinomial formula is written as

$$\frac{n!}{r_1!r_2!\dots r_l!} = \binom{n}{r_1}\binom{n-r_1}{r_2}\binom{n-r_1-r_2}{r_3}\cdots\binom{n-r_1-r_2-\dots-r_{l-2}}{r_{l-1}}. \tag{1.8-5}$$

To affect a realization of r_1 elements in the first subpopulation, r_2 in the second, and so on, we would select r_1 elements from the given n, r_2 from the remaining $n - r_1$, r_3 from the remaining $n - r_1 - r_2$, etc. But there are $\binom{n}{r_1}$ ways of choosing r_1 elements out of n, $\binom{n-r_1}{r_2}$ ways of choosing r_2 elements out of the remaining $n - r_1$, and so

forth. Thus, the total number of ways of choosing r_1 from n, r_2 from $n - r_1$, and so on is simply the product of the factors on the right-hand side of Equation 1.8-5 and the proof is complete. ■

Example 1.8-1

(*toss 12 dice*) [1-8, p. 36] Suppose we throw 12 dice; since each die throw has six outcomes there are a total of $n_T = 6^{12}$ outcomes. Consider now the event E that each face appears twice. There are, of course, many ways in which this can happen. Two outcomes in which this happens are shown below:

Dice I.D. Number	1	2	3	4	5	6	7	8	9	10	11	12
Outcome 1	3	1	3	6	1	2	5	4	4	6	2	5
Outcome 2	6	1	3	2	6	3	4	5	5	1	4	2

The total number of ways that this event can occur is the number of ways 12 dice ($n = 12$) can be arranged into six groups ($k = 6$) of two each ($r_1 = r_2 = \ldots = r_6 = 2$). Assuming that all outcomes are equally likely we compute

$$P[E] = \frac{n_E}{n_T} = \frac{\text{number of ways } E \text{ can occur}}{\text{total number of outcomes}}$$
$$= \frac{12!}{(2!)^6 6^{12}} = 0.003438.$$

The binomial and multinomial coefficients appear in the binomial and multinomial probability laws discussed in the next sections. The multinomial coefficient is also important in a class of problems called *occupancy problems* that occur in theoretical physics.

Occupancy Problems*

Occupancy problems are generically modeled as the random placement of r balls into n cells. For the first ball there are n choices, for the second ball there are n choices, and so on, so that there are n^r possible distributions of r balls in n cells and each has a probability of n^{-r}. If the balls are distinguishable, then each of the distributions is distinguishable; if the balls are not distinguishable, then there are fewer than n^r distinguishable distributions. For example, with three distinguishable balls ($r = 3$) labeled "1," "2," "3" and two cells ($n = 2$), we get eight (2^3) distinguishable distributions:

Cell no. 1	1	2	3	1,2	1,3	2,3	1,2,3	—
Cell no. 2	2,3	1,3	1,2	3	2	1	—	1,2,3

When the balls are not distinguishable (each ball is represented by a"*"), we obtain four distinct distributions:

Cell no. 1	***	**	*	—
Cell no. 2	—	*	**	***

How many distinguishable distributions can be formed from r balls and n cells? An elegant way to compute this is furnished by William Feller [1-8, p. 38] using a clever artifice. This artifice consists of representing the n cells by the spaces between $n+1$ bars and the balls by stars. Thus,

| | | |

represents three empty cells, while

|**| | |*|**| |*****|

represents two balls in the first cell, zero balls in the second and third cells, one in the fourth, two in the fifth, and so on. Indeed, with $r_i \geq 0$ representing the number of balls in the ith cell and r being the total number of balls, it follows that

$$r_1 + r_2 \ldots r_n = r.$$

The n-tuple $(r_1, r_2, \ldots, r_n)$ is called the *occupancy* and the r_i are the *occupancy numbers*; two distributions of balls in cells are said to be indistinguishable if their corresponding occupancies are identical. The occupancy of

|**| | |*|**| |*****|

is (2,0,0,1,2,0,5). Note that n cells require $n+1$ bars but since the first and last symbols must be bars, only $n-1$ bars and r stars can appear in any order. Thus, we are asking for the number of subpopulations of size r in a population of size $n-1+r$. The result is, by Equation 1.8-2,

$$\binom{n+r-1}{r} = \binom{n+r-1}{n-1}. \tag{1.8-6}$$

Example 1.8-2

(*distinguishable distributions*) Show that the number of distinguishable distributions in which no cell remains empty is $\binom{r-1}{n-1}$. Here we require that no bars be adjacent. Therefore, n of the r stars must occupy spaces between the bars but the remaining $r-n$ stars can go anywhere. Thus, $n-1$ bars and $r-n$ stars can appear in any order. The number of distinct distributions is then equal to the number of ways of choosing $r-n$ places in $(n-1)$ bars $+(r-n)$ stars or $r-n$ out of $n-1+r-n = r-1$. This is, by Equation 1.8-2,

$$\binom{r-1}{r-n} = \binom{r-1}{n-1}.$$

Example 1.8-3

(*birthdays on same date*) Small groups of people are amazed to find that their birthdays often coincide with others in the group. Before declaring this a mystery of fate, we analyze this situation as an occupancy problem. We want to compute how large a group is needed to have a high probability of a birthday *collision*, that is, at least two people in the group having their birthdays on the same date.

Solution We let the n ($n = 365$) days of the year be represented by n cells, and the r people in the group be represented by r balls. Then when a ball is placed into a cell, it fixes the birthday of the person represented by that ball. A birthday collision occurs when two or more balls are in the same cell. Now consider the arrangements of the balls. The first ball can go into any of the n cells, but the second ball has only $n-1$ cells to choose from to avoid a collision. Likewise, the third ball has only $n-2$ cells to choose from if a collision is to be avoided. Continuing in this fashion, we find that the number of arrangements that avoid a collision is $n(n-1)\cdots(n-r+1)$. The total number of arrangements of r balls in n cells is n^r. Hence with $P_0(r,n)$ denoting the probability of zero birthday collisions as a function of r and n, we find that $P_0(r,n) = \frac{n(n-1)\cdots(n-r+1)}{n^r} = \prod_{i=1}^{r-1}(1-\frac{i}{n})$. Then $1-P_0(r,n)$ is the probability of at least one collision.

How large does r need to be so $1 - P_0(r,n) \geq 0.9$ or, equivalently, $P_0(r,n) \leq 0.1$? Except for the mechanics of solving for r in $\prod_{i=1}^{r-1}(1-\frac{i}{n}) \leq 0.1$, the problem is over. We use a result from elementary calculus that for any real x, $1 - x \leq e^{-x}$, which is quite a good approximation for x near 0. If we replace $\prod_{i=1}^{r-1}(1-\frac{i}{n}) \leq 0.1$ by $\prod_{i=1}^{r-1} e^{-\frac{i}{n}} \leq 0.1$, we get a bound and estimate of r. Since $\prod_{i=1}^{r-1} e^{-\frac{i}{n}} = \exp\{-\frac{1}{n}\sum_{i=1}^{r-1} i\}$ and with use of $\sum_{i=1}^{r-1} i = r(r-1)/2$, it follows that $e^{-\frac{1}{2n}r(r-1)} \leq 0.1$ will give us an estimate of r. Solving for r and assuming that $r^2 >> r$, and $n = 365$, we get that $r \approx 40$. So having 40 people in a group will yield a 90 percent of (at least) two people having their birthdays on the same day.

Example 1.8-4

(*treize*) In seventeenth-century Venice, during the holiday of *Carnivale*, gamblers wearing the masks of the characters in the *commedia dell'arte* played the card game *treize* in entertainment houses called *ridottos.*

In *treize*, one player acts as the bank and the other players place their bets. The bank shuffles the deck, cards face down, and then calls out the names of the cards in order, from 1 to 13—ace to king—as he turns over one card at a time. If the card that is turned over matches the number he calls, then he (the bank) wins and collects all the bets. If the card that is turned over does not match the bank's call, the game continues until the dealer calls "thirteen." If the thirteenth card turned over is not a king, the bank loses the game and must pay each of the bettors an amount equal to their bet; in that case the player acting as bank must relinquish his position as bank to the player on the right.

What is the probability that the bank wins?

Solution We simplify the analysis by assuming that once a card is turned over, and there is no match, it is put back into the deck and the deck is reshuffled before the next card is dealt, that is, turned over. Let A_n denote the event that the bank has a *first* match, that is, a win, on the nth deal and W_n denote the event of a win in n tries. Since there are 4 cards of each number in a deck of 52 cards, the probability of a match is $1/13$. In order for a first win on the nth deal there have to be $n-1$ non matches followed by a match. The probability of this event is

$$P[A_n] = \left(1 - \frac{1}{13}\right)^{n-1} \frac{1}{13}.$$

Since $A_iA_j = \phi$ for $i \neq j$, the probability of a win in 13 tries is

$$P[W_{13}] = \sum_{i=1}^{13} P[A_i] = \frac{1}{13} \frac{1 - \left(\frac{12}{13}\right)^{13}}{\left(1 - \frac{12}{13}\right)} = 0.647,$$

from which it follows that the probability of the event W_{13}^c that the bank loses is $P[W_{13}^c] = 0.353$. Actually this result could have been more easily obtained by observing that the bank loses if it fails to get a match (no successes) in 13 tries, with success probability $1/13$. Hence

$$P[W_{13}^c] = \binom{13}{0}\left(\frac{1}{13}\right)^0\left(\frac{12}{13}\right)^{13} = 0.353.$$

Note that in the second equation we used the sum of the geometric series result: $\sum_{n=0}^{N-1} x^n = \frac{1-x^N}{1-x}$ (cf. Appendix A).

Points to consider. Why does $P[A_n] \to 0$ as $n \to \infty$? Why is $P[W_n] \geq P[A_n]$ for all n? How would you remodel this problem if we didn't make the assumption that the dealt card was put back into the deck? How would the problem change if the bank called down from 13 (king) to 1 (ace)?

In statistical mechanics, a six-dimensional space called *phase space* is defined as a space which consists of three position and three momentum coordinates. Because of the uncertainty principle which states that the uncertainty in position times the uncertainty in momentum cannot be less than Planck's constant h, phase space is quantized into tiny cells of volumes $v = h^3$. In a system that contains atomic or molecular size particles, the distribution of these particles among the cells constitutes the *state* of the system. In Maxwell–Boltzmann statistics, all distributions of r particles among n cells are equally likely. It can be shown (see, for example, *Concepts of Modern Physics* by A. Beiser, McGraw-Hill, 1973) that this leads to the famous Boltzmann law

$$n(\varepsilon) = \frac{2\pi N}{(\pi k)^{3/2}}\sqrt{\varepsilon}\, e^{-\varepsilon/kT}, \tag{1.8-7}$$

where $n(\varepsilon)d\varepsilon$ is the number of particles with energy between ε and $\varepsilon + d\varepsilon$, N is the total number of particles, T is absolute temperature, and k is the Boltzmann constant. The Maxwell–Boltzmann law holds for identical particles that, in some sense, can be distinguished. It is argued that the molecules of a gas are particles of this kind. It is not difficult to show that Equation 1.8-7 integrates to N.

In contrast to the Maxwell–Boltzmann statistics, where all n^r arrangements are equally likely, Bose–Einstein statistics considers only distinguishable arrangements of indistinguishable identical particles. For n cells and r particles, the number of such arrangements is given by Equation 1.8-6

$$\binom{n+r-1}{r},$$

and each arrangement is assigned a probability

$$\binom{n+r-1}{r}^{-1}.$$

It is argued that Bose–Einstein statistics are valid for photons, nuclei, and particles of zero or integral spin that do not obey the exclusion principle. The exclusion principle, discovered by Wolfgang Pauli in 1925, states that for a certain class of particles (e.g., electrons) no two particles can exist in the same quantum states (e.g., no two or more balls in the same cell).

To deal with particles that obey the exclusion principle, a third assignment of probabilities is construed. This assignment, called Fermi–Dirac statistics, assumes

(1) the exclusion principle (no two or more balls in the same cell); and
(2) all distinguishable arrangements satisfying (1) are equally probable.

Note that for Fermi–Dirac statistics, $r \leq n$. The number of distinguishable arrangements under the hypothesis of the exclusion principle is the number of subpopulations of size $r \leq n$ in a population of n elements or $\binom{n}{r}$. Since each is equally likely, the probability of any one state is $\binom{n}{r}^{-1}$.

The above discussions should convince the reader of the tremendous importance of probability in the basic sciences as well as its limitations: No amount of pure reasoning based on probability axioms could have determined which particles obey which probability laws.

Extensions and Applications

Theorem 1.5-1 on the probability of a union of events can be used to solve problems of engineering interest. First we note that the *number* of individual probability terms in the sum S_i is $\binom{n}{i}$. Why? There are a total of n indices and in S_i, all terms have i indices. For example, with $n = 5$ and $i = 2$, S_2 will consist of the sum of the terms P_{ij}, where the indices ij are 12; 13; 14; 15; 23; 24; 25; 34; 35; 45. Each set of indices in S_i never repeats, that is, they are all different. Thus, the number of indices and, therefore, the number of terms in S_i is the number of subpopulations of size i in a population of size n which is $\binom{n}{i}$ from Equation 1.8-2. Note that S_n will have only a single term.

Example 1.8-5

We are given r balls and n cells. The balls are indistinguishable and are to be randomly distributed among the n cells. Assuming that each arrangement is equally likely, compute the probability that all cells are occupied. Note that the balls may represent data packets and the cells buffers. Or, the balls may represent air-dropped food rations and the cells, people in a country in famine.

Solution Let E_i denote the event that cell i is empty ($i = 1, \ldots, n$). Then the r balls are placed among the remaining $n - 1$ cells. For each of the r balls there are $n - 1$ cells to

choose from. Hence there are $A(r, n-1) \triangleq (n-1)^r$ ways of arranging the r balls among the $n-1$ cells. Obviously, since the balls are indistinguishable, not all arrangements will be distinguishable. Indeed there are only $\binom{n+r-1}{n-1}$ distinguishable distributions and these are not, typically, equally likely. The total number of ways of distributing the r balls among the n cells is n^r. Hence

$$P[E_i] = (n-1)^r/n^r = \left(1 - \frac{1}{n}\right)^r \triangleq P_i.$$

Next assume that cells i and j are empty. Then $A(r, n-2) = (n-2)^r$ and

$$P[E_iE_j] \triangleq P_{ij} = \left(1 - \frac{2}{n}\right)^r.$$

In a similar fashion, it is easy to show that $P[E_iE_jE_k] = \left(1 - \frac{3}{n}\right)^r \triangleq P_{ijk}$, and so on. Note that the right-hand side expressions for P_i, P_{ij}, P_{ijk}, and so on do not contain the subscripts i, ij, ijk, and so on. Thus, each S_i contains $\binom{n}{i}$ identical terms and their sum amounts to

$$S_i = \binom{n}{i}\left(1 - \frac{i}{n}\right)^r.$$

Let E denote the event that at least one cell is empty. Then by Theorem 1.5-1,

$$P[E] = P\left[\bigcup_{i=1}^{n} E_i\right] = S_1 - S_2 + \ldots + S_n$$

Substituting for S_i from two lines above, we get

$$P[E] = \sum_{i=1}^{n}\binom{n}{i}(-1)^{i+1}\left(1 - \frac{i}{n}\right)^r. \tag{1.8-8}$$

The event that all cells are occupied is E^c. Hence $P[E^c] = 1 - P[E]$, which can be written as

$$P[E^c] = \sum_{i=0}^{n}\binom{n}{i}(-1)^{i}\left(1 - \frac{i}{n}\right)^r. \tag{1.8-9}$$

Example 1.8-6

(*m cells empty*) Use Equation 1.8-9 to compute the probability that exactly m out of the n cells are empty after the r balls have been distributed. We denote this probability by the three-parameter function $P_m(r,n)$.

Solution We write $P[E^c] = P_0(r,n)$. Now assume that exactly m cells are empty and $n-m$ cells are occupied. Next, let's fix the m cells that are empty, for example, cells numbers $\underbrace{2,4,5,7,\ldots,l}_{m \text{ terms}}$. Let $B(r,n-m)$ be the number of ways of distributing r balls among the remaining $n-m$ cells such that no cell remains empty and let $A(r,n-m)$ denote the number of ways of distributing r balls among $n-m$ cells. Then $P_0(r,n-m) = B(r,n-m)/A(r,n-m)$ and, since $A(r,n-m) = (n-m)^r$, we get that $B(r,n-m) = (n-m)^r P_0(r,n-m)$. There are $\binom{n}{m}$ ways of placing m empty cells among n cells. Hence the total number of arrangements of r balls among n cells such that m remain empty is $\binom{n}{m}(n-m)^r P_0(r,n-m)$. Finally, the number of ways of distributing r balls among n cells is n^r. Thus,

$$P_m(r,n) = \binom{n}{m}(n-m)^r P_0(r,n-m)/n^r.$$

or, after simplifying,

$$P_m(r,n) = \binom{n}{m}\sum_{i=0}^{n-m}\binom{n-m}{i}(-1)^i\left(1-\frac{i+m}{n}\right)^r. \tag{1.8-10}$$

1.9 BERNOULLI TRIALS—BINOMIAL AND MULTINOMIAL PROBABILITY LAWS

Consider the very simple experiment consisting of a single trial with a binary outcome: a success $\{\zeta_1 = \mathrm{s}\}$ with probability p, $0 < p < 1$, or a failure $\{\zeta_2 = \mathrm{f}\}$ with probability $q = 1-p$. Thus, $P[\mathrm{s}] = p$, $P[\mathrm{f}] = q$ and the sample space is $\Omega = \{\mathrm{s}, \mathrm{f}\}$. The σ-field of events $\mathscr{F}$ is ϕ, Ω, $\{\mathrm{s}\}$, $\{\mathrm{f}\}$. Such an experiment is called a *Bernoulli trial.*

Suppose we do the experiment twice. The new sample space Ω_2, written $\Omega_2 = \Omega \times \Omega$, is the set of all ordered 2-tuples

$$\Omega_2 = \{\mathrm{ss}, \mathrm{sf}, \mathrm{fs}, \mathrm{ff}\}.$$

$\mathscr{F}$ contains $2^4 = 16$ events. Some are ϕ, Ω, $\{\mathrm{ss}\}$, $\{\mathrm{ss}, \mathrm{ff}\}$, and so forth.

In the general case of n Bernoulli trials, the Cartesian product sample space becomes

$$\Omega_n = \underbrace{\Omega \times \Omega \times \ldots \times \Omega}_{n \text{ times}}$$

and contains 2^n elementary outcomes, each of which is an ordered n-tuple. Thus,

$$\Omega_n = \{a_1, \ldots, a_M\},$$

where $M = 2^n$ and $a_i \triangleq z_{i_1} \ldots z_{i_n}$, an ordered n-tuple, where z_{i_j} =s or f. Since each outcome z_{i_j} is independent of any other outcome, the joint probability $P[z_{i_1} \ldots z_{i_n}] = P[z_{i_1}]P[z_{i_2}] \ldots P[z_{i_n}]$. Thus, the probability of a given ordered set of k successes and $n-k$ failures is simply $p^k q^{n-k}$.

Example 1.9-1

(*repeated trials of coin toss*) suppose we throw a coin three times with $p = P[\text{H}]$ and $q = P[\text{T}]$. The probability of the event {HTH} is $pqp = p^2q$. The probability of the event {THH} is also p^2q. The different events leading to two heads and one tail are listed here:

$$E_1 = \{\text{HHT}\},$$

$$E_2 = \{\text{HTH}\},$$

$$E_3 = \{\text{THH}\}.$$

If F denotes the event of getting two heads and one tail without regard to order, then $F = E_1 \cup E_2 \cup E_3$. Since $E_iE_j = \phi$ for all $i \neq j$, we obtain $P[F] = P[E_1] + P[E_2] + P[E_3] = 3p^2q$.

Let us now generalize the previous result by considering an experiment consisting of n Bernoulli trials. The sample space Ω_n contains $M = 2^n$ outcomes $a_1, a_2, \ldots, a_M$, where each a_i is a string of n symbols, and each symbol represents a success s or a failure f. Consider the event $A_k \triangleq \{k \text{ successes in } n \text{ trials}\}$ and let the primed outcomes, that is, a_i', denote strings with k successes and $n-k$ failures. Then, with K denoting the number of ordered arrangements involving k successes and $n-k$ failures, we write

$$A_k = \bigcup_{i=1}^{K} \{a_i'\}.$$

To determine how large K is, we use an artifice similar to that used in proving Equation 1.8-6. Here, let represent failures and stars represent successes. Then, as an example,

$$| * * | * * | | *$$

represents five successes in nine tries in the order fssfssffs. How many such arrangements are there? The solution is given by Equation 1.8-6 with $r = k$ and $(n-1) + r$ replaced by $(n-k) + k = n$. (Note that there is no restriction that the first and last symbols must be bars.) Thus,

$$K = \binom{n}{k}$$

and, since $\{a_i\}$ are disjoint, that is, $\{a_i\} \cap \{a_j\} = \phi$ for all $i \neq j$, we obtain

$$P[A_k] = P\left[\bigcup_{i=1}^{K}\{a_i'\}\right] = \sum_{i=1}^{K} P[a_i'].$$

Finally, since $P[a_i'] = p^k q^{n-k}$ regardless of the ordering of the s's and f's, we obtain

$$P[A_k] = \binom{n}{k} p^k q^{n-k}$$
$$\triangleq b(k; n, p). \tag{1.9-1}$$

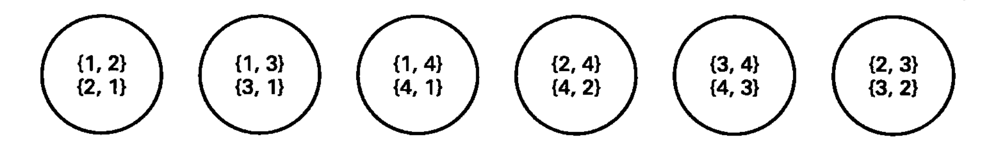
Binomial probability law. The three-parameter function $b(k; n, p)$ defined in Equation 1.9-1 is called the *binomial probability law* and is the probability of getting k successes in n independent tries with individual Bernoulli trial success probability p. The binomial coefficient

$$C_k^n = \binom{n}{k}$$

was introduced in the previous section and is the number of subpopulations of size k that can be formed from a population of size n. In Example 1.9-1 about tossing a coin three times, the population has size 3 (three tries) and the subpopulation has size 2 (two heads), and we were interested in getting two heads in three tries with order being irrelevant. Thus, the correct result is $C_2^3 = 3$. Note that had we asked for the probability of getting two heads on the first two tosses followed by a tail, that is, $P[E_1]$, we would not have used the coefficient C_2^3 since there is only one way that this event can happen.

Example 1.9-2

(*draw two balls from urn*) Suppose $n = 4$; that is, there are four balls numbered 1 to 4 in the urn. The number of distinguishable, ordered samples of size 2 that can be drawn without replacement is 12, that is, $\{1,2\}$; $\{1,3\}$; $\{1,4\}$; $\{2,1\}$; $\{2,3\}$; $\{2,4\}$; $\{3,1\}$; $\{3,2\}$; $\{3,4\}$; $\{4,1\}$; $\{4,2\}$; $\{4,3\}$. The number of distinguishable unordered sets is 6, that is,



From Equation 1.8-2 we obtain this result directly; that is ($n = 4$, $k = 2$)

$$\binom{n}{k} = \frac{4!}{2!2!} = 6.$$

Example 1.9-3

(*binary pulses*) Ten independent binary pulses per second arrive at a receiver. The error (i.e., a zero received as a one or vice versa) probability is 0.001. What is the probability of at least one error/second?

Solution

$$P[\text{at least one error/sec}] = 1 - P[\text{no errors/sec}]$$

$$= 1 - \binom{10}{0}(0.001)^0(0.999)^{10} = 1 - (0.999)^{10} \simeq 0.01.$$

Observation. Note that

$$\sum_{k=0}^{n} b(k; n, p) = 1. \quad \text{Why?}$$

Example 1.9-4

(*odd-man out*) An odd number of people want to play a game that requires two teams made up of even numbers of players. To decide who shall be left out to act as umpire, each of the N persons tosses a fair coin with the following stipulation: If there is one person whose outcome (be it heads or tails) is different from the rest of the group, that person will be the umpire. Assume that there are 11 players. What is the probability that a player will be "odd-man out," that is, will be the umpire on the first play?

Solution Let $E \triangleq \{10H, 1T\}$, where $10H$ means $H, H, \ldots, H$ ten times, and $F \triangleq \{10T, 1H\}$. Then $EF = \phi$ and

$$\begin{aligned} P[E \cup F] &= P[E] + P[F] \\ &= \binom{11}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right) + \binom{11}{1}\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{10} \\ &\simeq 0.01074. \end{aligned}$$

Example 1.9-5

(more *odd-man out*) In Example 1.9-4 derive a formula for the probability that the odd-man out will occur for the first time on the nth play. (*Hint*: Consider each play as an independent Bernoulli trial with success if an odd-man out occurs and failure otherwise.)

Solution Let E be the event of odd-man out for first time on the nth play. Let F be the event of no odd-man out in $n - 1$ plays and let G be the event of an odd-man out on the nth play. Then

$$E = FG.$$

Since it is completely reasonable to assume F and G are independent events, we can write

$$P[E] = P[F]P[G]$$

$$P[F] = \binom{n-1}{0}(0.0107)^0(0.9893)^{n-1} = (0.9893)^{n-1}$$

$$P[G] = 0.0107.$$

Thus, $P[E] = (0.0107)(0.9893)^{n-1}, n \geq 1$, which is often referred to as a *geometric distribution*† or law.

Example 1.9-6

(*multiple lottery tickets*) If you want to buy 50 lottery tickets, should you purchase them all in one lottery, or should you buy single tickets in 50 similar lotteries? For simplicity we take the case of a 100 ticket lottery with ticket prices of $1 each, and 50 such independent lotteries are available. Consider first the case of buying 50 tickets from one such lottery. Let E_i denote the event that the ith ticket is the winning ticket. Since any ticket is as likely to be the winning ticket as any other ticket, and not more than one ticket can be a winner, we have by classical probability that $P[E_i] = n_{win}/n_{tot} = 1/100$ for 1= 1,...,100. The event of winning the lottery is that one of the 50 purchased tickets is the winning ticket or, equivalently, with E denoting the event that one of the 50 tickets is the winner $E = \cup_{i=1}^{50} E_i$ and $P[E] = P[\cup_{i=1}^{50} E_i] = \sum_{i=1}^{50} P[E_i] = 50 \times 1/100 = 0.5$. Next we consider the case of buying 1 ticket in each of 50 separate lotteries. We recognize this as Bernoulli trials with an individual success probability $p = 0.01$ and $q = 0.99$. With the aid of a calculator, we can find the probability of winning (exactly) once as

$$\begin{aligned} P[\text{win once}] &= b(1; 50, 0.01) \\ &= \binom{50}{1}(0.01)^1(0.99)^{49} \\ &= 50 \times 10^{-2} \times 0.611 \\ &\doteq 0.306^{\ddagger} \end{aligned}$$

Similarly, we find the probability of winning twice $b(2; 50, 0.01) \doteq 0.076$, the probability of winning three times $b(3; 50, 0.01) \doteq 0.012$, the probability of winning four times $b(4; 50, 0.01) \doteq 0.001$, and the probability of winning more times is negligible. As a check we can easily calculate the probability of winning at least once,

$$\begin{aligned} P[\text{win at least once}] &= 1 - P[\text{loose every time}] \\ &= 1 - q^{50} \\ &= 1 - (0.99)^{50} \\ &\doteq 0.395. \end{aligned}$$

†A popular variant on this definition is the alternative geometric distribution given as $pq^n, n \geq 0$ with $q = 1 - p$ and $0 \leq p \leq 1$.

‡We use the notation [equals sign with dot over top] to indicate that all the decimal digits shown are correct.

Indeed we have $0.395 = 0.306+0.076+0.012+0.001$. We thus find that, if your only concern is to win at least once, it is better to buy all 50 tickets from one lottery. On the other hand, when playing in separate lotteries, there is the possibility of winning multiple times. So your average winnings may be more of a concern. Assuming a fair lottery with payoff \$100, we can calculate an average winnings as

$$
\begin{aligned}
&100 \times 0.306 + 200 \times 0.076 + 300 \times 0.012 + 400 \times 0.001 \\
&\doteq 49.8.
\end{aligned}
$$

So, in terms of average winnings, it is about the same either way.

Further discussion of the binomial law. We write down some formulas for further use. The probability $B(k; n, p)$ of k or fewer *successes in n tries* is given by

$$B(k; n, p) = \sum_{i=0}^{k} b(i; n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i q^{n-i}. \tag{1.9-2}$$

The symbol $B(k; n, p)$ is called the *binomial distribution function.* The probability of k or more successes in n tries is

$$\sum_{i=k}^{n} b(i; n, p) = 1 - B(k-1; n, p).$$

The probability of more than k successes but no more than j successes is

$$\sum_{i=k+1}^{j} b(i; n, p) = B(j; n, p) - B(k; n, p).$$

There will be much more on distribution functions in later Chapters. We illustrate the application of this formula in Example 1.9-7.

Example 1.9-7

(*missile attack*) Five missiles are fired against an aircraft carrier in the ocean. It takes at least two direct hits to sink the carrier. All five missiles are on the correct trajectory but must get through the "point-defense" guns of the carrier. It is known that the point-defense guns can destroy a missile with probability $p = 0.9$. What is the probability that the carrier will still be afloat after the encounter?

Solution Let E be the event that the carrier is still afloat and let F be the event of a missile getting through the point-defense guns. Then

$$P[F] = 0.1$$

and

$$P[E] = 1 - P[E^c]$$
$$= 1 - \sum_{i=2}^{5} \binom{5}{i} (0.1)^i (0.9)^{5-i} \simeq 0.92.$$

Multinomial Probability Law

The multinomial probability law is a generalization of the binomial law. The binomial law is based on Bernoulli trials in which only two outcomes are possible. The multinomial law is based on a generalized Bernoulli trial in which l outcomes are possible. Thus, consider an elementary experiment consisting of a single trial with k elementary outcomes $\zeta_1, \zeta_2, \ldots, \zeta_l$. Let the probability of outcome ζ_i be p_i $(i = 1, \ldots, l)$. Then

$$p_i \geq 0, \quad \text{and} \sum_{i=1}^{l} p_i = 1. \tag{1.9-3}$$

Assume that this generalized Bernoulli trial is repeated n times and consider the event consisting of a *prescribed, ordered* string of elementary outcomes in which ζ_1 appears r_1 times, ζ_2 appears r_2 times, and so on until ζ_l appears r_l times. What is the probability of this event? The key here is that the order is prescribed *a priori*. For example, with $l = 3$ (three possible outcomes) and $n = 6$ (six tries), a prescribed string might be $\zeta_1\zeta_3\zeta_2\zeta_2\zeta_1\zeta_2$ so that $r_1 = 2$, $r_2 = 3$, $r_3 = 1$. Observe that $\sum_{i=1}^{l} r_i = n$. Since the outcome of each trial is an independent event, the probability of observing a prescribed ordered string is $p_1^{r_1} p_2^{r_2} \ldots p_l^{r_l}$. Thus, for the string $\zeta_1\zeta_3\zeta_2\zeta_2\zeta_1\zeta_2$ the probability is $p_1^2 p_2^3 p_3$.

A different (greater) probability results when order is not specified. Suppose we perform n repetitions of a generalized Bernoulli trial and consider the event in which ζ_1 appears r_1 times, ζ_2 appears r_2 times, and so forth, *without regard to order.* Before computing the probability of this event we furnish an example.

Example 1.9-8

(*busy emergency number*) In calling the Sav-Yur-Life health care facility to report an emergency, one of three things can happen:

(1) the line is busy (event E_1);
(2) you get the wrong number (event E_2); and
(3) you get through to the triage nurse (event E_3).

Assume $P[E_i] = p_i$. What is the probability that in five separate emergencies at different times, initial calls are met with four busy signals and one wrong number?

Solution Let F denote the event of getting four busy signals and one wrong number. Then

$$F = F_1 \cup F_2 \cup F_3 \cup F_4 \cup F_5, \text{ where } F_1 = \{E_1E_1E_1E_1E_2\},$$

$$F_2 = \{E_1E_1E_1E_2E_1\},\ F_3 = \{E_1E_1E_2E_1E_1\},\ F_4 = \{E_1E_2E_1E_1E_1\},$$

and

$$F_5 = \{E_2E_1E_1E_1E_1\}.$$

Since $F_iF_j = \phi$, $P[F] = \sum_{i=1}^{5} P[F_i]$. But $P[F_i] = p_1^4 p_2^1 p_3^0$ independent of i. Hence

$$P[F] = 5p_1^4 p_2^1 p_3^0.$$

With the assumed $p_1 = 0.3$, $p_2 = 0.1$, $p_3 = 0.6$, we get

$$P[F] = 5 \times 8.1 \times 10^{-3} \times 0.1 \times 1 = 0.004.$$

In problems of this type we must count all the strings of length n in which ζ_1 appears r_1 times, ζ_2 appears r_2 times, and so on. In the example just considered, there were five such strings. In the general case of n trials with r_1 outcomes of ζ_1, r_2 outcomes of ζ_2, and so on, there are

$$\frac{n!}{r_1!r_2!\dots r_l!}, \tag{1.9-4}$$

such strings. In Example 1.9-8, $n = 5$, $r_1 = 4$, $r_2 = 1$, $r_3 = 0$ so that

$$\frac{5!}{4!1!0!} = 5.$$

The number in Equation 1.9-4 is recognized as the multinomial coefficient. To check that it is the appropriate coefficient, consider the r_1 outcomes ζ_1. The number of ways of placing the r_1 outcomes ζ_1 among the n trials is identical with the number of subpopulations of size r_1 in a population of size n which is $\binom{n}{r_1}$. That leaves $n - r_1$ trials among which we wish to place r_2 outcomes ζ_2. The number of ways of doing that is $\binom{n-r_1}{r_2}$. Repeating this process we obtain the total number of distinguishable arrangements

$$\binom{n}{r_1}\binom{n-r_1}{r_2}\dots\binom{n-r_1-r_2\dots-r_{l-1}}{r_l} = \frac{n!}{r_1!r_2!\dots r_l!}$$

Example 1.9-9

(*repeated generalized Bernoulli*) Consider four repetitions of a generalized Bernoulli experiment in which the outcomes are $*$, $\bullet$, 0. What is the number of ways of getting two $*$, one $\bullet$, and one 0.

Solution The number of ways of getting two $*$ in four trials is $\binom{4}{2} = 6$. If we let the spaces between bars represent a trial, then we can denote the outcomes as

$$|*|*|\ |\ |;\ |*|\ |*|\ |;\ |\ |*|*|\ |;\ |\ |*|\ |*|;\ |\ |\ |*|*|;\ |*|\ |\ |*|.$$

The number of ways of placing • among the two remaining cells is $\binom{2}{1} = 2$. The number of ways of placing 0 among the remaining cell is $\binom{1}{1} = 1$. Hence the total number of arrangements is $6 \cdot 2 \cdot 1 = 12$. They are

*	0	•	*
*	•	0	*
*	*	0	•
*	*	•	0
*	0	*	•
*	•	*	0
0	*	*	•
•	*	*	0
0	*	•	*
•	*	0	*
0	•	*	*
•	0	*	*

We can now state the multinomial probability law. Consider a generalized Bernoulli trial with outcomes $\zeta_1, \zeta_2, \ldots, \zeta_l$ and let the probability of observing outcome ζ_i be p_i, $i = 1, \ldots, l$, where $p_i \geq 0$ and $\sum_{i=1}^{l} p_i = 1$. The probability that in n trials ζ_1 occurs r_1 times, ζ_2 occurs r_2 times, and so on is

$$P(\boldsymbol{r}; n, \boldsymbol{p}) = \frac{n!}{r_1! r_2! \ldots r_l!} p_1^{r_1} p_2^{r_2} \ldots p_l^{r_l}, \tag{1.9-5}$$

where $\boldsymbol{r}$ and $\boldsymbol{p}$ are l-tuples defined by

$$\boldsymbol{r} = (r_1, r_2, \ldots, r_l), \quad \boldsymbol{p} = (p_1, p_2, \ldots, p_l), \text{ and } \sum_{i=1}^{l} r_i = n.$$

Observation. With $l = 2$, Equation 1.9-5 becomes the binomial law with $p_1 \triangleq p$, $p_2 \triangleq 1 - p$, $r_1 \triangleq k$, and $r_2 \triangleq n - k$. Functions such as Equations 1.9-1 and 1.9-5 are often called *probability mass functions.*

Example 1.9-10

(*emergency calls*) In the United States, 911 is the all-purpose number used to summon an ambulance, the police, or the fire department. In the rowdy city of Nirvana in upstate New York, it has been found that 60 percent of all calls request the police, 25 percent request an ambulance, and 15 percent request the fire department. We observe the next ten calls. What is the probability of the combined event that six calls will ask for the police, three for ambulances, and one for the fire department?

Solution Using Equation 1.9-5 we get

$$P(6,3,1;10,0.6,0.25,0.15)$$
$$= \frac{10!}{6!3!1!}(0.6)^6(0.25)^3(0.15)^1 \simeq 0.092.$$

A numerical problem appears if n gets large. For example, suppose we observe 100 calls and consider the event of 60 calls for the police, 30 for ambulances, and 10 for the fire department; clearly computing numbers such as 100!, 60!, 30! requires some care. An important result that helps in evaluating such large factorials is Stirling's[†] formula:

$$n! \simeq (2\pi)^{1/2} n^{n+(1/2)} e^{-n},$$

where the approximation improves as n increases, for example,

n	$n!$	Stirling's formula	Percent error
1	1	0.922137	8
10	3,628,800	3,598,700	0.8

When using a numerical computer to evaluate Equation 1.9-5, additional care must be used to avoid loss of accuracy due to under- and over-flow. A joint evaluation of pairs of large and small numbers can help in this regard, as can the use of logarithms.

As stated earlier, the binomial law is a special case, perhaps the most important case, of the multinomial law. When the parameters of the binomial law attain extreme values, the binomial law can be used to generate another important probability law. This is explored next.

1.10 ASYMPTOTIC BEHAVIOR OF THE BINOMIAL LAW: THE POISSON LAW

Suppose that in the binomial function $b(k;n,p)$, $n >> 1$, $p << 1$, but np remains constant, say $np = \mu$. Recall that $q = 1 - p$. Hence

$$\binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{1}{k!}\mu^k \left(1 - \frac{\mu}{n}\right)^{n-k},$$

where $n(n-1)\ldots(n-k+1) \simeq n^k$ if n is allowed to become large enough and k is held fixed. Hence in the limit as $n \to \infty$, $p \to 0$, and $k << n$, we obtain

$$b(k;n,p) \simeq \frac{1}{k!}\mu^k \left(1 - \frac{\mu}{n}\right)^{n-k} \xrightarrow{n\to\infty} \frac{\mu^k}{k!} e^{-\mu}. \tag{1.10-1}$$

[†] James Stirling, eighteenth-century mathematician.

Thus, in situations where the binomial law applies with $n >> 1$, $p << 1$ but $np = \mu$ is a finite constant, we can use the approximation

$$b(k; n, p) \simeq \frac{\mu^k}{k!} e^{-\mu}. \tag{1.10-2}$$

Poisson law. The *Poisson probability law,* with parameter $\mu(> 0)$, is given as

$$p(k) = \frac{\mu^k}{k!} e^{-\mu}, \quad 0 \leq k < \infty.$$

Unlike the binomial law, the Poisson law just has one parameter μ that can take on any positive value.

Example 1.10-1

(*time to failure*) A computer contains 10,000 components. Each component fails independently from the others and the yearly failure probability per component is 10^{-4}. What is the probability that the computer will be working one year after turn-on? Assume that the computer fails if one or more components fail.

Solution

$$p = 10^{-4}, \qquad n = 10{,}000, \qquad k = 0, \qquad np = 1.$$

Hence

$$b(0; 10{,}000, 10^{-4}) = \frac{1^0}{0!} e^{-1} = \frac{1}{e} = 0.368.$$

Example 1.10-2

(*random points in time*) Suppose that n independent points are placed at random in an interval $(0, T)$. Let $0 < t_1 < t_2 < T$ and $t_2 - t_1 \triangleq \tau$. Let $\tau/T << 1$ and $n >> 1$. What is the probability of observing exactly k points in τ seconds? (Figure 1.10-1.)

Solution Consider a single point placed at random in $(0, T)$. The probability of the point appearing in τ is τ/T. Let $p = \tau/T$. Every other point has the same probability of being in τ seconds. Hence, the probability of finding k points in τ seconds is the binomial law

$$P[k \text{ points in } \tau \text{ sec}] = \binom{n}{k} p^k q^{n-k}. \tag{1.10-3}$$

With $n >> 1$, we use the approximation in Equation 1.10-1 to give

$$b(k; n, p) \simeq \left(\frac{n\tau}{T}\right)^k \frac{e^{-(n\tau/T)}}{k!}, \tag{1.10-4}$$

where n/T can be interpreted as the "average" number of points per unit interval.

Replacing the average rate in this example with parameter μ ($\mu > 0$), we get the *Poisson* law defined by

$$P[k \text{ points}] = e^{-\mu} \frac{\mu^k}{k!}, \tag{1.10-5}$$

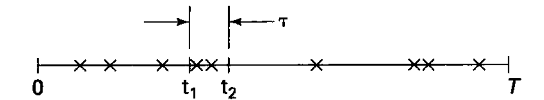


Figure 1.10-1 Points placed at random on a line. Each point is placed with equal likelihood anywhere along the line.

where $k = 0, 1, 2, \ldots$. With $\mu \triangleq \lambda\tau$, where λ is the average number of points† per unit time and τ is the length of the interval $(t, t+\tau]$, the probability of k points in elapsed time τ is

$$P(k; t, t+\tau) = e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!}. \tag{1.10-6}$$

For the Poisson law, we also stipulate that *numbers of points arriving in disjoint time intervals constitute independent events.* We can regard this as inherited from an underlying set of Bernoulli trials, which are always independent.

In Equation 1.10-6 we assume that λ is a constant and not a function of t. If λ varies with t, we can generalize $\lambda\tau$ with the integral $\int_t^{t+\tau} \lambda(u)\, du$, and the probability of k points in the interval $(t,\ t+\tau]$ becomes

$$P(k; t, t+\tau) = \exp\left[-\int_t^{t+\tau} \lambda(u)\, du\right] \frac{1}{k!} \left[\int_t^{t+\tau} \lambda(u) du\right]^k. \tag{1.10-7}$$

The Poisson law $P[k$ events in $\Delta x]$, or more generally $P[k$ events in $(x, x+\Delta x)]$, where x is time, volume, distance, and so forth and Δx is the interval associated with x, is widely used in engineering and sciences. Some typical situations in various fields where the Poisson law is applied are listed below.

Physics. In radioactive decay—$P[k\,\alpha$-particles in τ seconds] with λ the average number of emitted α-particles per second.

Engineering. In planning the size of a call center—$P[k$ telephone calls in τ seconds] with λ the average number of calls per second.

Biology. In water pollution monitoring—$P[k$ coliform bacteria in 1000 cubic centimeters] with λ the average number of coliform bacteria per cubic centimeter.

Transportation. In planning the size of a highway toll facility—$P[k$ automobiles arriving in τ minutes] with λ the average number of automobiles per minute.

Optics. In designing an optical receiver—$P[k$ photons per second over a surface of area $A]$ with λ the average number of photons-per-second per unit area.

Communications. In designing a fiber optical transmitter–receiver link—$P[k$ photoelectrons generated at the receiver in one second] with λ the average number of photoelectrons per second.

†The term *points* here is a generic term. Equally appropriate would be "arrivals," "hits," "occurrences," etc.

The parameter λ is often called, in this context, the Poisson *rate parameter*. Its dimensions are points per unit interval, the interval being time, distance, volume, and so forth. When the form of the Poisson law that we wish to use is as in Equation 1.10-6 or 1.10-7, we speak of the Poisson law with *rate parameter* λ or *rate function* $\lambda(t)$.

Example 1.10-3

(*misuse of probability*) (a) "Prove" that there must be life in the universe, other than that on our own Earth, by using the following numbers[†]: average number of stars per galaxy, 300×10^9; number of galaxies, 100×10^9; probability that a star has a planetary system, 0.5; average number of planets per planetary system, 9; probability that a planet can sustain life, 1/9; probability, p, of life emerging on a life-sustaining planet, 10^{-12}.

Solution First we compute, n_{LS}, the number of planets that are life-sustaining:

$$\begin{aligned} n_{\text{LS}} &= 300 \times 10^9 \times 100 \times 10^9 \times 0.5 \times 9 \times 1/9 \\ &= 1.5 \times 10^{22}. \end{aligned}$$

Next we use the Poisson approximation to the binomial with $a = n_{\text{LS}}\, p = 1.5 \times 10^{22} \times 10^{-12}$, for computing the probability of *no life outside of Earth's* and obtain

$$b(0, 1.5 \times 10^{22}, 10^{-12}) = \frac{(1.5 \times 10^{10})^0}{0!} e^{-1.5\times 10^{10}} \approx 0.$$

Hence we have just "shown" that the probability of life outside Earth has a probability of unity, that is, a sure bet. Note that the number for life emerging on other planets, 10^{-12}, is impressively low.

(b) Now show that life outside Earth is extremely unlikely by using the same set of numbers except that the probability of life emerging on a life-sustaining planet has been reduced to 10^{-30}.

Solution Using the Poisson approximation to the binomial, with $a = 1.5 \times 10^{22} \times 10^{-30} = 1.5 \times 10^{-8}$, we obtain for the probability of no life outside Earth's:

$$\begin{aligned} b(0, 1.5 \times 10^{22}, 10^{-30}) &= \frac{(1.5 \times 10^{-8})^0}{0!} e^{-1.5\times 10^{-8}} \\ &\approx 1 - (1.5 \times 10^{-8}) \approx 1, \end{aligned}$$

where we have used the approximation $e^{-x} \approx 1 - x$ for small x.

Thus, by changing only one number, we have gone from "proving" that the universe contains extraterrestrial life to proving that, outside of ourselves, the universe is lifeless. The reason that this is a misuse of probability is that, at present, we have no idea as to the factors that lead to the emergence of life from nonliving material. While the calculation is technically correct, this example illustrates the use of contrived numbers to either prove or disprove what is essentially a belief or faith.

[†] All the numbers have been quoted at various times by proponents of the idea of extraterrestrial life.

Example 1.10-4

(*website server*) A website server receives on the average 16 access requests per minute. If the server can handle at most 24 accesses per minute, what is the probability that in any one minute the website will saturate?

Solution Saturation occurs if the number of requests in a minute exceeds 24. The probability of this event is

$$P[\text{saturation}] = \sum_{k=25}^{\infty} [\lambda\tau]^k \frac{e^{-\lambda\tau}}{k!} \tag{1.10-8}$$

$$= \sum_{k=25}^{\infty} [16]^k \frac{e^{-16}}{k!} \simeq 0.017 \simeq 1/60. \tag{1.10-9}$$

Thus, about once in every 60 minutes (on the "average") will a visitor be turned away.

Given the numerous applications of the Poisson law in engineering and the sciences, one would think that its origin is of somewhat more noble birth than "merely" as a limiting form of the binomial law. Indeed this is the case, and the Poisson law can be derived once three assumptions are made. Obviously these three assumptions should reasonably mirror the characteristics of the underlying physical process; otherwise our results will be of only marginal interest. Fortunately, in many situations these assumptions seem to be quite valid.

In order to be concrete, we shall talk about occurrences taking place in *time* (as opposed to, say, length or distance). The Poisson law is based on the following three assumptions:

1. The probability, $P(1; t, t+\Delta t)$, of a single event occurring in $(t, t+\Delta t]$ is proportional to Δt, that is,
$$P(1; t, t+\Delta t) \simeq \lambda(t)\Delta t \qquad \Delta t \to 0. \tag{1.10-10}$$
In Equation 1.10-10, $\lambda(t)$ is the Poisson rate parameter.
2. The probability of k $(k > 1)$ events in $(t, t+\Delta t]$ goes to zero:
$$P(k; t, t+\Delta t) \simeq 0 \qquad \Delta t \to 0, \qquad k = 2, 3, \ldots. \tag{1.10-11}$$
3. Events in nonoverlapping time intervals are statistically independent.†

Starting with these three simple physical assumptions, it is a straightforward task to obtain the Poisson probability law. We leave this derivation to Chapter 9 but merely point out that the clever use of the assumptions leads to a set of elementary, first-order differential equations whose solution is the Poisson law. The general solution is furnished by Equation 1.10-7 but, fortunately, in a large number of physical situations the Poisson rate

†Note in property 3 we are talking about disjoint time intervals, not disjoint events. For disjoint events we would add probabilities, but for disjoint time intervals which lead to independent events in the Poisson law, we multiply the individual probabilities.

parameter $\lambda(t)$ can be approximated by a constant, say, λ. In that case Equation 1.10-6 can be applied. We conclude this section with a final example.

Example 1.10-5

(*defects in digital tape*) A manufacturer of computer tape finds that the defect density along the length of tape is not uniform. After a careful compilation of data, it is found that for tape strips of length D, the defect density $\lambda(x)$ along the tape length x varies as

$$\lambda(x) = \lambda_0 + \frac{1}{2}(\lambda_1 - \lambda_0)\left(1 + \cos\left[\frac{2\pi x}{D}\right]\right), \qquad \lambda_1 > \lambda_0$$

for $0 \leq x \leq D$ due to greater tape contamination at the edges $x = 0$ and $x = D$.

(a) What is the meaning of $\lambda(x)$ in this case?
(b) What is the average number of defects for a tape strip of length D?
(c) What is an expression for k defects on a tape strip of length D?
(d) What are the Poisson assumptions in the case?

Solution

(a) Bearing in mind that $\lambda(x)$ is a *defect density*, that is, the average number of defects per unit length at x, we conclude that $\lambda(x)\Delta x$ is the average number of defects in the tape from x to $x + \Delta x$.

(b) Given the definition of $\lambda(x)$, we conclude that the average number of defects along the whole tape is merely the integral of $\lambda(x)$, that is,

$$\begin{aligned}\int_0^D \lambda(x)dx &= \int_0^D \left[\lambda_0 + \frac{1}{2}(\lambda_1 - \lambda_0)\left(1 + \cos\frac{2\pi x}{D}\right)\right] dx \\ &= \frac{\lambda_0 + \lambda_1}{2} D \\ &\triangleq \Lambda.\end{aligned}$$

(c) Assuming the Poisson law holds, we use Equation 1.10-7 with x and Δx (distances) replacing t and τ (times). Thus,

$$P(k; x, x + \Delta x) = \exp\left[-\int_x^{x+\Delta x} \lambda(\zeta)d\zeta\right] \cdot \frac{1}{k!}\left[\int_x^{x+\Delta x} \lambda(\zeta)d\zeta\right]^k.$$

In particular, with $x = 0$ and $x + \Delta x = D$, we obtain

$$P(k; 0, D) = \Lambda^k \frac{e^{-\Lambda}}{k!},$$

where Λ is as defined above.

(d) The Poisson assumptions become

(i) $P[1; x, x + \Delta x] \simeq \lambda(x)\Delta x, \quad \text{as} \quad \Delta x \to 0.$

(ii) $P[k; x, x+\Delta x] = 0 \quad \Delta x \to 0, \quad \text{for} \quad k = 2, 3, \ldots$; that is, the probability of there being more than one defect in the interval $(x, x + \Delta x)$ as Δx becomes vanishingly small is zero.

(iii) the occurrences of defects (events) in nonoverlapping sections of the tape are independent.

1.11 NORMAL APPROXIMATION TO THE BINOMIAL LAW

In this section we give, without proof, a numerical approximation to binomial probabilities and binomial sums. Let S_k denote the event consisting of (exactly) k successes in n Bernoulli trials. Then the probability of S_k follows a binomial distribution and

$$P[S_k] = \binom{n}{k} p^k q^{n-k} = b(k; n, p), \quad 0 \le k \le n. \tag{1.11-1}$$

For large values of n and k, Equation 1.11-1 may be difficult to evaluate numerically. Also, the probability of the event $\{k_1 <\text{number of successes} \le k_2\}$ may involve many terms, making a direct evaluation of its probability $P[k_1 <\text{number of successes} \le k_2]$ difficult. Fortunately, when n is large, we can use approximate methods for evaluating such probabilities. These approximate methods involve the so-called *Normal or Gaussian distribution.*

The Normal distribution and its significance will be discussed in greater detail in Chapter 2 and subsequent chapters in this book. Here we use it only to help evaluate binomial probabilities. For the present, define the function $f_{SN}(x)$, known as the *standard Normal density*, by

$$f_{SN}(x) \triangleq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \tag{1.11-2}$$

and its running integral, known as the *standard Normal cumulative distribution function*, by

$$F_{SN}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}y^2\right) dy. \tag{1.11-3}$$

Then, when n is large it can be shown [1-8] that

$$b(k; n, p) \approx \frac{1}{\sqrt{npq}} f_{SN}\left(\frac{k - np}{\sqrt{npq}}\right). \tag{1.11-4}$$

The approximation becomes better when $npq >> 1$. We reproduce the results from [1-8] in Table 1.11-1. Even in this case, $npq = 1.6$, the approximation is quite good. The approximation for sums, when $n >> 1$ and k_1 and k_2 are fixed integers, takes the form

$$P[k_1 < \text{number of successes} \le k_2] \approx F_{SN}\left[\frac{k_2 - np + 0.5}{\sqrt{npq}}\right] - F_{SN}\left[\frac{k_1 - np - 0.5}{\sqrt{npq}}\right]. \tag{1.11-5}$$

Table 1.11-1 Normal Approximation to the Binomial for Selected Numbers

k	$b(k; 10, 0.2)$	Normal approximation
0	0.1074	0.0904
1	0.2864	0.2307
2	0.3020	0.3154
3	0.2013	0.2307
4	0.0880	0.0904
5	0.0264	0.0189
6	0.0055	0.0021

Table 1.11-2 Event Probabilities Using the Normal Approximation (Adapted from [1-8])

n	p	α	β	$P[\alpha \leq S_n \leq \beta]$	Normal approximation
200	0.5	95	105	0.5632	0.5633
500	0.1	50	55	0.3176	0.3235
100	0.3	12	14	0.00015	0.00033
100	0.3	27	29	0.2379	0.2341
100	0.3	49	51	0.00005	0.00003

Some results, for various values of n, p, k_1, k_2, are furnished in Table 1.11-2, which uses the results in [1-8].

In using the Normal approximation, one should refer to Table 2.4-1. In Table 2.4-1 a function called $\text{erf}(x)$ is given rather than $F_{SN}(x)$. The $\text{erf}(x)$ is defined by

$$\text{erf}(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{y^2}{2}} dy.$$

However, since it is easy to show that

$$F_{SN}(x) = \tfrac{1}{2} + \text{erf}(x), \quad x > 0, \tag{1.11-6}$$

and

$$F_{SN}(x) = \tfrac{1}{2} - \text{erf}(|x|), \quad x < 0, \tag{1.11-7}$$

we can compute Equation 1.11-5 in terms of the table values. Thus, with $a \triangleq \frac{k_1 - np - 0.5}{\sqrt{npq}}$ and $b \triangleq \frac{k_2 - np + 0.5}{\sqrt{npq}}$ and $b' > a'$, we can use the results in Table 1.11-3.

The Normal approximation is also useful in evaluating Poisson sums. For example, a sum such as in Equation 1.10-9 is tedious to evaluate if done directly. However, if $\lambda\tau >> 1$, we can use the *Normal approximation to the Poisson law*, which is merely an extension of the Normal approximation to the binomial law. This extension is expected since we have seen that the Poisson law is itself an approximation to the binomial law under certain circumstances. From the results given above we are able to justify the following approximation.

$$\sum_{k=\alpha}^{\beta} e^{-\lambda\tau} \frac{[\lambda\tau]^k}{k!} = \frac{1}{\sqrt{2\pi}} \int_{l_1}^{l_2} \exp\left(-\frac{1}{2}y^2\right) dy, \tag{1.11-8}$$

where

$$l_2 \triangleq \frac{\beta - \lambda\tau + 0.5}{\sqrt{\lambda\tau}}$$

and

$$l_1 \triangleq \frac{\alpha - \lambda\tau - 0.5}{\sqrt{\lambda\tau}}.$$

Another useful approximation is

$$e^{-\lambda\tau} \frac{[\lambda\tau]^k}{k!} = \frac{1}{\sqrt{2\pi}} \int_{l_3}^{l_4} \exp\left(-\frac{1}{2}y^2\right) dy, \tag{1.11-9}$$

where

$$l_4 \triangleq \frac{k - \lambda\tau + 0.5}{\sqrt{\lambda\tau}}$$

and

$$l_3 \triangleq \frac{k - \lambda\tau - 0.5}{\sqrt{\lambda\tau}}.$$

For example, with $\lambda\tau = 5$, and $k = 5$, the error in using the Normal approximation of Equation 1.11-9 is less than 1 percent.

SUMMARY

In this, the first chapter of the book, we have reviewed some different definitions of probability. We developed the axiomatic theory and showed that for a random experiment three important objects were required: the sample space Ω, the sigma field of events $\mathscr{F}$, and a probability measure P. The mathematical triple $(\Omega, \mathscr{F}, P)$ is called the probability space $\mathscr{P}$.

We introduced the important notions of independent, dependent, and compound events, and conditional probability. We developed a number of relations to enable the application of these.

We discussed some important formulas from combinatorics and briefly illustrated how important they were in theoretical physics. We then discussed the binomial probability law and its generalization, the multinomial law. We saw that the binomial law could, when certain limiting conditions were valid, be approximated by the Poisson law. The Poisson law, one of the central laws in probability theory, was shown to have application in numerous branches of science and engineering. We stated, but deferred verification until Chapter 9, that the Poisson law can be derived directly from simple and entirely reasonable physical assumptions.

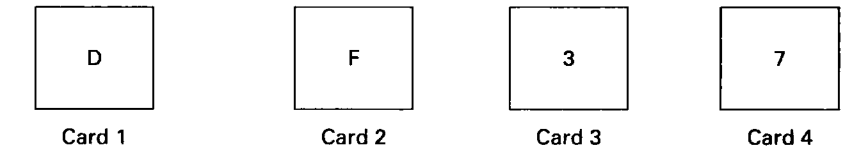
Approximations for the binomial and Poisson laws, based on the Normal distribution, were furnished. Several occupancy problems of engineering interest were discussed. In Chapter 4 we shall revisit these problems.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

1.1 In order for a statement such as "Ralph is probably guilty of theft" to have meaning in the relative frequency approach to probability, what kind of data would one need?

1.2 Problems in applied probability (a branch of mathematics called statistics) often involve testing P $\rightarrow Q$ (P implies Q) type statements, for example, if she smokes, she will probably get sick; if he is smart he will do well in school. You are given a set of four cards that have a letter on one side and a number on the other. You are asked to test the rule "If a card has a D on one side, it has a three on the other." Which of the following cards should you turn over to test the veracity of the rule:



Card 1 Card 2 Card 3 Card 4

Be careful here!

1.3 In a spinning-wheel game, the spinning wheel contains the numbers 1 to 9. The contestant wins if an *odd* number shows. What is the probability of a win? What are your assumptions?

1.4 A fair coin is flipped three times. The outcomes on each flip are heads H or tails T. What is the probability of obtaining two tails and one head?

1.5 An urn contains three balls numbered 1, 2, 3. The experiment consists of drawing a ball at random, recording the number, and replacing the ball before the next ball is drawn. This is called sampling with replacement. What is the probability of drawing the same ball thrice in three tries?

1.6 An experiment consists of drawing two balls *without* replacement from an urn containing five balls numbered 1 to 5. Describe the sample space Ω. What is Ω if the ball is replaced before the second is drawn?

1.7 The experiment consists of measuring the heights of each partner of a randomly chosen married couple. (a) Describe Ω in convenient notation; (b) let E be the event that the man is shorter than the woman. Describe E in convenient notation.

1.8 An urn contains ten balls numbered 1 to 10. Let E be the event of drawing a ball numbered no greater than 6. Let F be the event of drawing a ball numbered greater than 3 but less than 9. Evaluate E^c, F^c, EF, $E \cup F$, EF^c, $E^c \cup F^c$, $EF^c \cup E^cF$, $EF \cup E^cF^c$, $(E \cup F)^c$, and $(EF)^c$. Express these events in words.

1.9 There are four *equally likely* outcomes $\zeta_1, \zeta_2, \zeta_3$, and ζ_4 and two events $A = \{\zeta_1, \zeta_2\}$ and $B = \{\zeta_2, \zeta_3\}$. Express the sets (events) AB^c, BA^c, AB, and $A \cup B$ in terms of their elements (outcomes).

1.10 Verify the useful set identities $A = AB \cup AB^c$ and $A \cup B = (AB^c) \cup (BA^c) \cup (AB)$. Does probability add over these unions? Why?

1.11 In a given random experiment there are four *equally likely* outcomes $\zeta_1, \zeta_2, \zeta_3$, and ζ_4. Let the event $A \triangleq \{\zeta_1, \zeta_2\}$. What is the probability of A? What is the event (set) A^c in terms of the outcomes? What is the probability of A^c? Verify that $P[A] = 1 - P[A^c]$ here.

1.12 Consider the probability space $(\Omega, \mathcal{F}, P)$ for this problem.

(a) State the three axioms of probability theory and explain in a sentence the significance of each.

(b) Derive the following formula, justifying each step by reference to the appropriate axiom above,

$$P[A \cup B] = P[A] + P[B] - P[A \cap B],$$

where A and B are arbitrary events in the field $\mathcal{F}$.

1.13 An experiment consists of drawing two balls at random, with replacement from an urn containing five balls numbered 1 to 5. Three students "Dim," "Dense," and "Smart" were asked to compute the probability p that the sum of numbers appearing on the two draws equals 5. Dim computed $p = \frac{2}{15}$, arguing that there are 15 distinguishable unordered pairs and only 2 are favorable, that is, $(1, 4)$ and $(2, 3)$. Dense computed $p = \frac{1}{9}$, arguing that there are 9 distinguishable sums (2 to 10), of which only 1 was favorable. Smart computed $p = \frac{4}{25}$, arguing that there were 25 distinguishable ordered outcomes of which 4 were favorable, that is, $(4, 1)$, $(3, 2)$, $(2, 3)$, and $(1, 4)$. Why is $p = \frac{4}{25}$ the correct answer? Explain what is wrong with the reasoning of Dense and Dim.

1.14 Prove the distributive law for set union, that is,

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

by showing that each side is contained in the other.

1.15 Prove the general result $P[A] = 1 - P[A^c]$ for any probability experiment and any event A defined on this experiment.

1.16 Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. Define three events: $A = \{1, 2\}$, $B = \{2, 3\}$, and $C = \{4, 5, 6\}$. The probability measure is unknown, but it satisfies the three axioms.

(a) What is the probability of $A \cap C$?
(b) What is the probability of $A \cup B \cup C$?
(c) State a condition on the probability of either B or C that would allow them to be independent events.

1.17 Use the axioms given in Equations 1.5-1 to 1.5-3 to show the following: ($E \in \mathscr{F}$, $F \in \mathscr{F}$) (a) $P[\phi] = 0$; (b) $P[EF^c] = P[E] - P[EF]$; (c) $P[E] = 1 - P[E^c]$.

1.18 Use the probability space $(\Omega, \mathcal{F}, P)$ for this problem. What is the difference between an *outcome*, an *event*, and a *field of events*?

1.19 Use the axioms of probability to show the following: ($A \in \mathcal{F}, B \in \mathcal{F}$): $P[A \cup B] = P[A] + P[B] - P[A \cap B]$, where P is the probability measure on the sample space Ω, and $\mathcal{F}$ is the field of events.

1.20 Use the "exclusive-or" operator in Equation 1.4-3 to show that $P[E \oplus F] = P[EF^c] + P[E^cF]$.

1.21 Show that $P[E \oplus F]$ in the previous problem can be written as $P[E \oplus F] = P[E] + P[F] - 2P[EF]$.

***1.22** Let the sample space $\Omega = \{\text{cat, dog, goat, pig}\}$.

(a) Assume that only the following probability information is given:

$$P[\{\text{cat, dog}\}] = 0.9,$$
$$P[\{\text{goat, pig}\}] = 0.1,$$
$$P[\{\text{pig}\}] = 0.05,$$
$$P[\{\text{dog}\}] = 0.5.$$

For this given set of probabilities, find the appropriate field of events $\mathscr{F}$ so that the overall probability space $(\Omega, \mathscr{F}, P)$ is well defined. Specify the field $\mathscr{F}$ by listing all the events in the field, along with their corresponding probabilities.

(b) Repeat part (a), but without the information that $P[\{\text{pig}\}] = 0.05$.

1.23 Prove the distributive law for set intersection, that is,

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$

by showing that each side is contained in the other.

1.24 The probability that a communication system will have high fidelity is 0.81, and the probability that it will have high fidelity and high selectivity is 0.18. What is the probability that a system with high fidelity will also have high selectivity?

1.25 An urn contains eight balls. The letters a and b are used to label the balls. Two balls are labeled a, two are labeled b, and the remaining balls are labeled with both letters, that is a, b. Except for the labels, all the balls are identical. Now a ball is drawn at random from the urn. Let A and B represent the events of observing letters a and b, respectively. Find $P[A]$, $P[B]$, and $P[AB]$. Are A and B independent? (Note that you will observe the letter a when you draw either an a ball or an a, b ball.)

1.26 A fair die is tossed twice (a die is said to be fair if all outcomes $1, \ldots, 6$ are equally likely). Given that a 3 appears on the first toss, what is the probability of obtaining the *sum* 7 after the second toss?

1.27 In the experiment of throwing two fair dice, A is the event that the number on the first die is odd, B the event that the number on the second die is odd, and C the event that the sum of the faces is odd. Show that A, B and C are pairwise independent, but A, B and C are not independent.

1.28 Two numbers are chosen at random from the numbers 1 to 10 without replacement. Find the probability that the second number chosen is 5.

1.29 A random-number generator generates integers from 1 to 9 (inclusive). All outcomes are equally likely; each integer is generated independently of any previous integer. Let Σ denote the sum of two consecutively generated integers; that is, $\Sigma = N_1 + N_2$. Given that Σ is odd, what is the conditional probability that Σ is 7? Given that $\Sigma > 10$, what is the conditional probability that at least one of the integers is > 7? Given that $N_1 > 8$, what is the conditional probability that Σ will be odd?

1.30 Two firms, V and W, consider bidding on a road-building job which may or may not be awarded depending on the amount of the bids. Firm V submits a bid and the probability is 3/4 that V will get the job, provided firm W does not bid. The odds are 3 to 1 that W will bid and if it does, the probability that V will get the job is only 1/3.

(a) What is the probability that V will get the job?
(b) If V gets the job, what is the probability that W did not bid?

1.31 Henrietta is 29 years old and physically very fit. In college she majored in geology. During her student days, she frequently hiked in the national forests and biked in the national parks. She participated in anti-logging and anti-mining operations. Now, Henrietta works in an office building in downtown Nirvana. Which is greater: the probability that Henrietta's occupation is that of office manager; or the probability that Henrietta is an office manager who is active in nature-defense organizations like the Sierra Club?

1.32 In the ternary communication channel shown in Figure P1.32 a 3 is sent three times more frequently than a 1, and a 2 is sent two times more frequently than a 1. A 1 is observed; what is the conditional probability that a 1 was sent?

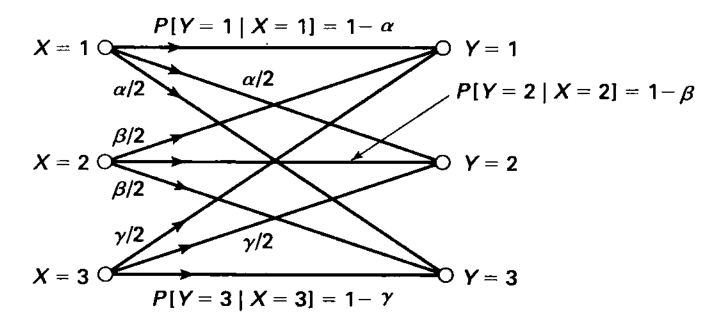


Figure P1.32 Ternary communication channel.

1.33 A large class in probability theory is taking a multiple-choice test. For a particular question on the test, the fraction of examinees who know the answer is p; $1-p$ is the fraction that will guess. The probability of answering a question correctly is unity for an examinee who knows the answer and $1/m$ for a guessee; m is the number of multiple-choice alternatives. Compute the probability that an examinee knew the answer to a question given that he or she has correctly answered it.

1.34 In the beauty-contest problem, Example 1.6-12, what is the probability of picking the most beautiful contestant if we decide *a priori* to choose the ith $(1 \leq i \leq N)$ contestant?

1.35 Assume there are three machines A, B, and C in a semiconductor manufacturing facility that make chips. They manufacture, respectively, 25, 35, and 40 percent of the total semiconductor chips there. Of their outputs, respectively, 6, 4, and 2 percent of the chips are defective. A chip is drawn randomly from the combined output of the three machines and is found defective. What is the probability that this defective chip was manufactured by machine A? by machine B? by machine C?

1.36 In Example 1.6-12, plot the probability of making a correct decision versus α/N, assuming that the "wait-and-see" strategy is adopted. In particular, what is $P[D]$ when $\alpha/N = 0.5$. What does this suggest about the sensitivity of $P[D]$ vis-a-vis α when α is not too far from α_0 and N is large?

1.37 In the village of Madre de la Paz in San Origami, a great flood displaces 103 villagers. The government builds a temporary tent village of 30 tents and assigns the 103 villagers randomly to the 30 tents.

(a) Identify this problem as an occupancy problem. What are the analogues to the balls and cells?
(b) How many distinguishable distributions of people in tents can be made?
(c) How many distinguishable distributions are there in which no tent remains empty?

***1.38** Consider r indistinguishable balls (particles) and n cells (states) where $n > r$. The r balls are placed at random into the n cells (multiple occupancy is possible). What is the probability P that the r balls appear in r preselected cells (one to a cell)?

***1.39** A committee of 5 people is to be selected randomly from a group of 5 men and 10 women. Find the probability that the committee consists of (a) 2 men and 3 women, and (b) only women.

***1.40** Three tribal elders win elections to lead the unstable region of North Vatisthisstan. Five identical assault rifles, a gift of the people of Sodabia, are airdropped among a meeting of the three leaders. The tribal leaders scamper to collect as many of the rifles as they each can carry, which is five.

(a) Identify this as an occupancy problem.
(b) List all possible distinguishable distribution of rifles among the three tribal leaders.
(c) How many distinguishable distributions are there where at least one of the tribal leaders fails to collect any rifles?
(d) What is the probability that all tribal leaders collect at least one rifle?

(e) What is the probability that exactly one tribal leader will not collect any rifles?

1.41 In some casinos there is the game *Sic bo*, in which bettors bet on the outcome of a throw of three dice. Many bets are possible each with a different payoff. We list some of them below with the associated payoffs in parentheses:

(a) Specified three of a kind (180 to 1);
(b) Unspecified three of a kind (30 to 1);
(c) Specified two of a kind (10 to 1);
(d) Sum of three dice equals 4 or 17 (60 to 1)
(e) Sum of three dice equals 5 or 16 (30 to 1);
(f) Sum of three dice equals 6 or 15 (17 to 1);
(g) Sum of three dice equals 7 or 14 (12 to 1)
(h) Sum of three dice equals 8 or 13 (8 to 1);
(i) Sum of three dice equals 9, 10, 11, 12 (6 to 1);
(j) Specified two dice combination; that is, of the three dice displayed, two of them must match exactly the combination wagered (5 to 1).

We wish to compute the associated probabilities of winning from the player's point of view and his expected gain.

1.42 Most communication networks use *packet switching* to create virtual circuits between two users, even though the users are sharing the same physical channel with others. In packet switching, the data stream is broken up into *packets* that travel different paths and are reassembled in the proper chronological order and at the correct address. Suppose the *order* information is missing. Compute the probability that a data stream broken up into N packets will reassemble itself correctly, even without the order information.

1.43 In the previous problem assume that $N = 4$. A lazy engineer decides to omit the order information in favor of repeatedly sending the data stream until the packets re-order correctly for the first time. Derive a formula that the correct re-ordering occurs for the first time on the nth try. How many repetitions should be allowed before the cumulative probability of a correct re-ordering for the first time is at least 0.95?

1.44 Prove that the binomial law $b(k; n, p)$ is a valid probability assignment by showing that $\sum_{k=0}^{n} b(k; n, p) = 1$.

1.45 War-game strategists make a living by solving problems of the following type. There are 6 incoming ballistic missiles (BMs) against which are fired 12 antimissile missiles (AMMs). The AMMs are fired so that two AMMs are directed against each BM. The single-shot-kill probability (SSKP) of an AMM is 0.8. The SSKP is simply the probability that an AMM destroys a BM. Assume that the AMM's don't interfere with each other and that an AMM can, at most, destroy only the BM against which it is fired. Compute the probability that (a) all BMs are destroyed, (b) at least one BM gets through to destroy the target, and (c) exactly one BM gets through.

1.46 Assume in the previous problem that the target was destroyed by the BMs. What is the conditional probability that only one BM got through?

1.47 A computer chip manufacturer finds that, historically, for every 100 chips produced, 80 meet specifications, 15 need reworking, and 5 need to be discarded. Ten chips are chosen for inspection.

(a) What is the probability that all 10 meet specs?
(b) What is the probability that 2 or more need to be discarded?
(c) What is the probability that 8 meet specs, 1 needs reworking, and 1 will be discarded?

1.48 Unlike the city of Nirvana, New York, where 911 is the all-purpose telephone number for emergencies, in Moscow, Russia, you dial 01 for a fire emergency, 02 for the police, and 03 for an ambulance. It is estimated that emergency calls in Russia have the same frequency distribution as in Nirvana, namely, 60 percent are for the police, 25 percent are for ambulance service, and 15 percent are for the fire department. Assume that 10 calls are monitored and that none of the calls overlap in time and that the calls constitute independent trials.

1.49 A smuggler, trying to pass himself off as a glass-bead importer, attempts to smuggle diamonds by mixing diamond beads among glass beads in the proportion of one diamond bead per 2000 beads. A harried customs inspector examines a sample of 100 beads. What is the probability that the smuggler will be caught, that is, that there will be at least one diamond bead in the sample?

1.50 Assume that a faulty receiver produces audible clicks to the great annoyance of the listener. The average number of clicks per second depends on the receiver temperature and is given by $\lambda(\tau) = 1 - e^{-\tau/10}$, where τ is time from turn-on. Evaluate the formula for the probability of $0, 1, 2, \ldots$ clicks during the first 5 seconds of operation after turn-on. Assume the Poisson law.

1.51 A frequently held lottery sells 100 tickets at \$1 per ticket every time it is held. One of the tickets must be a winner. A player has \$50 to spend. To maximize the probability of winning at least one lottery, should he buy 50 tickets in one lottery or one ticket in 50 lotteries?

1.52 In the previous problem, which of the two strategies will lead to a greater expected gain for the player? The expected gain if $M (M \leq 50)$ lotteries are played is defined as $\overline{G}_M \triangleq \sum_{i=1}^{M} G_i P(i)$, where G_i is the gain obtained in winning i lotteries.

1.53 The switch network shown in Figure P1.53 represents a digital communication link. Switches α_i $i = 1, \ldots, 6$, are open or closed and operate independently. The probability that a switch is closed is p. Let A_i represent the event that switch i is closed.

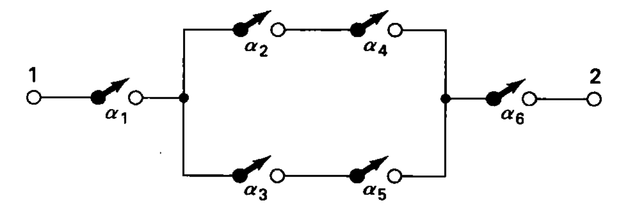


Figure P1.53 Switches in telephone link.

(a) In terms of the A_i's write the event that there exists at least one closed path from 1 to 2.
(b) Compute the probability of there being at least one closed path from 1 to 2.

1.54 (*independence of events in disjoint intervals for Poisson law*) The average number of cars arriving at a tollbooth per minute is λ and the probability of k cars in the interval $(0, T)$ minutes is

$$P(k; 0, T) = e^{-\lambda T}\frac{[\lambda T]^k}{k!}.$$

Consider two disjoint, that is, nonoverlapping, intervals, say $(0, t_1]$ and $(t_1, T]$. Then for the Poisson law:

$$P[n_1 \text{ cars in } (0, t_1] \text{ and } n_2 \text{ cars in } (t_1, T]] \tag{1.11-10}$$

$$= P[n_1 \text{ cars in } (0, t_1]]P[n_2 \text{ cars in } (t_1, T]], \tag{1.11-11}$$

that is events in disjoint intervals are independent. Using this fact, show the following:
(a) That $P[n_1 \text{ cars in } (0, t_1]|n_1 + n_2 \text{ cars in } (0, T]]$ is not a function of λ.
(b) In (a) let $T = 2$, $t_1 = 1$, $n_1 = 5$, and $n_2 = 5$. Compute $P[5 \text{ cars in } (0, 1]|10 \text{ cars in } (0, 2]$.

1.55 An automatic breathing apparatus (B) used in anesthesia fails with probability P_B. A failure means death to the patient unless a monitor system (M) detects the failure and alerts the physician. The monitor system fails with probability P_M. The failures of the system components are independent events. Professor X, an M.D. at Hevardi Medical School, argues that if $P_M > P_B$ installation of M is useless.† Show that Prof. X needs to take a course on probability theory by computing the probability of a patient dying with and without the monitor system in place. Take $P_M = 0.1 = 2P_B$.

1.56 In a particular communication network, the server broadcasts a packet of data (say, L bytes long) to N receivers. The server then waits to receive an acknowledgment message from each of the N receivers before proceeding to broadcast the next packet. If the server does not receive all the acknowledgments within a certain time period, it will rebroadcast (retransmit) the same packet. The server is then said to be in the "retransmission mode." It will continue retransmitting the packet until all N acknowledgments are received. Then it will proceed to broadcast the next packet. Let $p \triangleq P$[successful transmission of a single packet to a single receiver along with successful acknowledgment]. Assume that these events are independent for different receivers or separate transmission attempts. Due to random impairments in the transmission media and the variable condition of the receivers, we have that $p < 1$.

†A true story! The name of the medical school has been changed.

(a) In a fixed protocol or method of operation, we require that all N of the acknowledgments be received in response to a given transmission attempt for that packet transmission to be declared successful. Let the event $S(m)$ be defined as follows: $S(m) \triangleq$ {a successful transmission of one packet to all N receivers in m or fewer attempts}. Find the probability

$$P(m) \triangleq P[S(m)].$$

[*Hint*: Consider the complement of the event $S(m)$.]

(b) An improved system operates according to a dynamic protocol as follows. Here we relax the acknowledgment requirement on retransmission attempts, so as to only require acknowledgments from those receivers that have not yet been heard from on previous attempts to transmit the current packet. Let $S_D(m)$ be the same event as in part (a) but using the dynamic protocol. Find the probability

$$P_D(m) \triangleq P[S_D(m)].$$

[*Hint*: First consider the probability of the event $S_D(m)$ for an individual receiver, and then generalize to the N receivers.]

Note: If you try $p = 0.9$ and $N = 5$ you should find that $P(2) < P_D(2)$.

1.57 Toss two unbiased dice (each with six faces: 1 to 6), and write down the sum of the two face numbers. Repeat this procedure 100 times. What is the probability of getting 10 readings of value 7? What is the Poisson approximation for computing this probability? (*Hint*: Consider the event $A = \{sum = 7\}$ on a single toss and let p in Equation 1.9-1 be $P[A]$.)

1.58 On behalf of your tenants you have to provide a laundry facility. Your choices are

1. lease two inexpensive "Clogger" machines at \$50.00/month each; or
2. lease a single "NeverFail" at \$100/month.

The Clogger is out of commission 40 percent of the time while the NeverFail is out of commission only 20 percent of the time.

(a) From the tenant's point, which is the better alternative?
(b) From your point of view as landlord, which is the better alternative?

1.59 In the politically unstable country of Eastern Borduria, it is not uncommon to find a bomb onboard passenger aircraft. The probability that on any given flight, a bomb will be onboard is 10^{-2}. A nervous passenger always flies with an unarmed bomb in his suitcase, reasoning that the probability of there being *two bombs onboard* is 10^{-4}. By this maneuver, the nervous passenger believes that he has greatly reduced the airplane's chances of being blown up. Do you agree with his reasoning? If not, why not?

1.60 In a ring network consisting of eight links as shown in Figure P1.60, there are two paths connecting any two terminals. Assume that links fail independently with probability q, $0 < q < 1$. Find the probability of successful transmission of a packet

from terminal A to terminal B. (Note: Terminal A transmits the packet in *both* directions on the ring. Also, terminal B removes the packet from the ring upon reception. Successful transmission means that terminal B received the packet from either direction.)

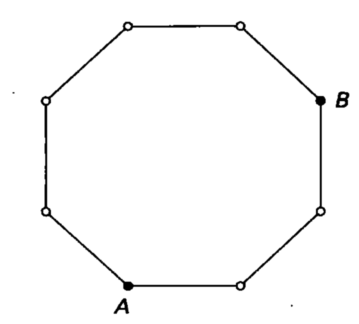


Figure P1.60 A ring network with eight stations.

1.61 A union directive to the executives of the telephone company demands that telephone operators receive overtime payment if they handle more than 5680 calls in an eight-hour day. What is the probability that Curtis, a unionized telephone operator, will collect overtime on a particular day where the occurrence of calls during the eight-hour day follows the Poisson law with rate parameter $\lambda = 710$ calls/hour?

1.62 Toss two unbiased coins (each with two sides: numbered 1 and 2), and write down the sum of the two side numbers. Repeat this procedure 80 times. What is the probability of getting 10 readings of value 2? What is the Poisson approximation for computing this probability?

1.63 The average number of cars arriving at a tollbooth is λ cars per minute, and the probability of cars arriving is assumed to follow the Poisson law. Given that 6 cars arrive in the first three minutes, what is the probability of 12 cars arriving in the first six minutes?

***1.64** An aging professor, desperate to finally get a good review for his course on probability, hands out chocolates to his students. The professor's short-term memory is so bad that he can't remember which students have already received a chocolate. Assume that, for all intents and purposes, the chocolates are distributed randomly. There are 10 students and 15 chocolates. What is the probability that each student received at least one chocolate?

1.65 Assume that code errors in a computer program occur as follows: A line of code contains errors with probability $p = 0.001$ and is error free with probability $q = 0.999$. Also errors in different lines occur independently. In a 1000-line program, what is the approximate probability of finding 2 or more erroneous lines?

1.66 Let us assume that two people have their birthdays on the same day if both the month and the day are the same for each (not necessarily the year). How many people would you need to have in a room before the probability is $\frac{1}{2}$ or greater that at least two people have their birthdays on the same day?

1.67 (*sampling*) We draw ten chips at random from a semiconductor manufacturing line that is known to have a defect rate of 2 percent. Find the probability that more than one of the chips in our sample is defective.

***1.68** (*percolating fractals*) Consider a square lattice with N^2 cells, that is, N cells per side. Write a program that does the following: With probability p you put an electrically conducting element in a cell and with probability $q = 1-p$, you leave the cell empty. Do this for every cell in the lattice. When you are done, does there exist a continuous path for current to flow from the bottom of the lattice to the top? If yes, the lattice is said to *percolate*. Percolation models are used in the study of epidemics, spread of forest fires, and *ad hoc* networks, etc. The lattice is called a *random fractal* because of certain invariant properties that it possesses. Try $N = 10, 20, 50$; $p = 0.1, 0.3, 0.6$. You will need a random number generator. MATLAB has the function *rand*, which generates uniformly distributed random numbers x_i in the interval $(0.0, 1.0)$. If the number $x_i \leq p$, make the cell electrically conducting; otherwise leave it alone. Repeat the procedure as often as time permits in order to estimate the probability of percolation for different p's. A nonpercolating lattice is shown in Figure P1.68(a); a percolating lattice is shown in (b). For more discussion of this problem, see M. Schroeder, *Fractals, Chaos, Power Laws* (New York: W.H. Freeman, 1991).

***1.69** You are a contestant on a TV game show. There are three identical closed doors leading to three rooms. Two of the rooms contain nothing, but the third contains a $100,000 Rexus luxury automobile which is yours if you pick the right door. You are asked to pick a door by the master of ceremonies (MC) who knows which room contains the Rexus. After you pick a door, the MC opens a door (not the one you picked) to show a room *not* containing the Rexus. Show that even without any further knowledge, you will greatly increase your chances of winning the Rexus if you *switch* your choice from the door you originally picked to the one remaining closed door.

1.70 Often we are faced with determining the *more likely* of two alternatives. In such a case we are given two probability measures for a single sample space and field of events, that is, $(\Omega, \mathcal{F}, P_1)$ and $(\Omega, \mathcal{F}, P_2)$, and we are asked to determine the probability of an observed event E in both cases. The *more likely* alternative is said to be the one which gives the higher probability of event E.

Consider that two coins are in a box; one is "fair" with $P_1[\{H\}] = 0.5$ and one is "biased" with $P_2[\{H\}] = p$. Without looking, we draw one coin from the box and then flip this single coin ten times. We only consider the repeated coin-flips as our experiment and so the sample space $\Omega = \{$ all ten-character strings of H and T$\}$. We observe the event $E = \{$a total of four H's and six T's$\}$.

(a) What are the two probabilities of the observed event E, that is, $P_1[E]$ and $P_2[E]$?

(b) Determine the *likelihood ratio* $L \triangleq P_1[E]/P_2[E]$ as a function of p. (When $L > 1$, we say that the fair coin is more likely. This test is called a *likelihood ratio test.*)

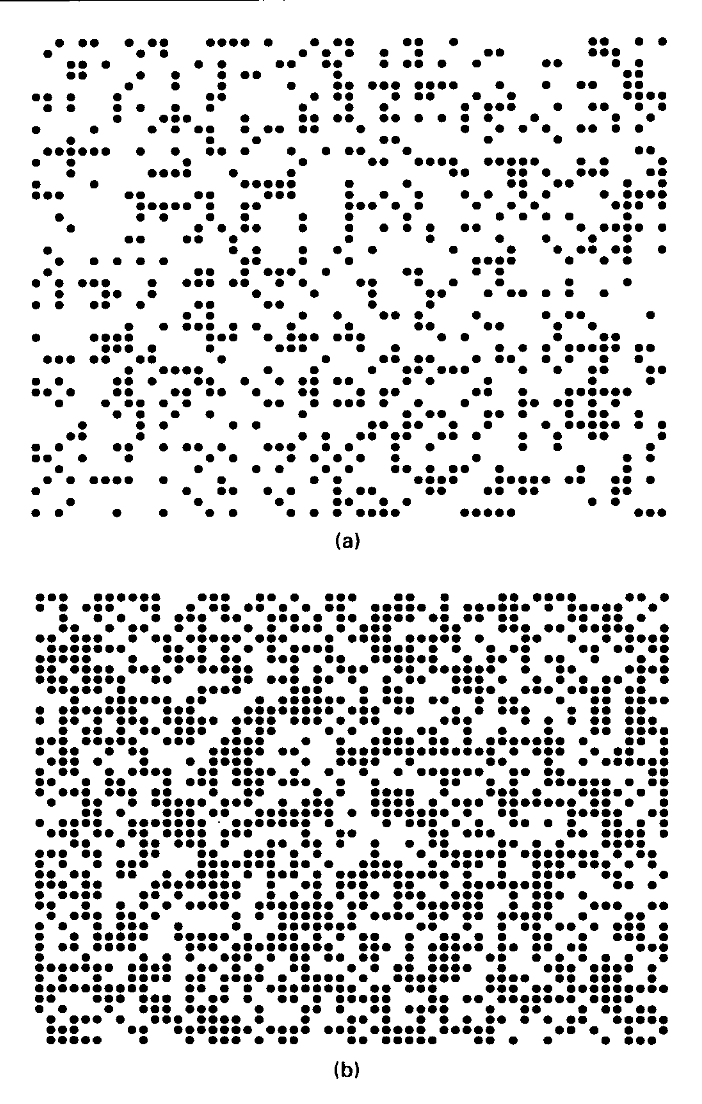


Figure P1.68 (a) Nonpercolating random fractal; (b) percolating random fractal. Can you find a percolating path? (From *Fractals, Chaos, Power Laws* by M. Schroeder, W.H. Freeman, New York, 1991. With permission.)

REFERENCES

1-1. E. Merzbacher, *Quantum Mechanics.* New York: John Wiley, 1961.
1-2. A. Kolmogorov, *Foundations of the Theory of Probability.* New York: Chelsea, 1950.
1-3. S. Pinker, *How the Mind Works.* New York: Norton, 1997.

1-4. B. O. Koopman, "The Axioms of Algebra and Intuitive Probability," *Annals of Mathematics* (2), Vol. 41, 1940, pp. 269–292.
1-5. A. Papoulis, and S. U. Pillai *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 4th Ed, 2002.
1-6. R. Von Mises, *Wahrscheinlichkeit, Statistic and Wahrheit.* Vienna: Springer-Verlag, 1936.
1-7. W. B. Davenport, Jr., *Probability and Random Processes.* New York: McGraw-Hill, 1970.
1-8. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd edition. New York: John Wiley, 1950, Chapter 2.
1-9. E. Parzen, *Modern Probability Theory and Its Applications.* New York: John Wiley, 1960, p. 119.

2 Random Variables

2.1 INTRODUCTION

Many random phenomena have outcomes that are sets of real numbers: the voltage $v(t)$, at time t, across a noisy resistor, the arrival time of the next customer at a movie theatre, the number of photons in a light pulse, the brightness level at a particular point on the TV screen, the number of times a light bulb will switch on before failing, the lifetime of a given living person, the number of people on a New York to Chicago train, and so forth. In all these cases the sample spaces are sets of numbers on the real line.

Even when a sample space Ω is not numerical, we might want to generate a new sample space from Ω that is numerical, that is, converting random speech, color, gray tone, and so forth to numbers, or converting the physical fitness profile of a person chosen at random into a numerical "fitness" vector consisting of weight, height, blood pressure, heart rate, and so on, or describing the condition of a patient afflicted with, say, black lung disease by a vector whose components are the number and size of lung lesions and the number of lung zones affected.

In science and engineering, we are in almost all instances interested in numerical outcomes, whether the underlying experiment $\mathscr{H}$ is numerical-valued or not. To obtain numerical outcomes, we need a rule or *mapping* from the original sample space Ω to the real line R^1. Such a mapping is what a random variable fundamentally is and we discuss it in some detail in the next several sections.

Let us, however, make a remark or two. The concept of a random variable will enable us to replace the original probability space with one in which events are sets of numbers.

Thus, on the induced probability space of a random variable every event is a subset of R^1. But is every subset of R^1 always an event? Are there subsets of R^1 that could get us into trouble via violating the axioms of probability? The answer is yes, but fortunately these subsets are not of engineering or scientific importance. We say that they are *nonmeasurable.*† Sets of practical importance are of the form $\{x = a\}$, $\{x\colon a \leq x \leq b\}$, $\{x\colon a < x \leq b\}$, $\{x\colon a \leq x < b\}$, $\{x\colon a < x < b\}$, and their unions and intersections. These five intervals are more easily denoted $[a]$, $[a, b]$, $(a, b]$, $[a, b)$, and (a, b). Intervals that include the end points are said to be *closed*; those that leave out end points are said to be *open*. Intervals can also be half-closed (half-open) too; for example, the interval $(a, b]$ is open on the left and closed on the right. The field of subsets of R^1 generated by the intervals was called the *Borel field* in Chapter 1, Section 4.

We can define more than one random variable on the same underlying sample space Ω. For example, suppose that Ω consists of a large, representational group of people in the United States. Let the experiment consist of choosing a person at random. Let X denote the person's lifetime and Y denote that person's daily consumption of cigarettes. We can now ask: Are X and Y related? That is, can we predict X from observing Y? Suppose we define a third random variable Z that denotes the person's weight. Is Z related to X or Y?

The main advantage of dealing with random variables is that we can define certain probability functions that make it both convenient and easy to compute the probabilities of various events. These functions must naturally be consistent with the axiomatic theory. For this reason we must be a little careful in defining events on the real line. Elaboration of the ideas introduced in this section is given next.

2.2 DEFINITION OF A RANDOM VARIABLE

Consider an experiment $\mathscr{H}$ with sample space Ω. The elements or points of Ω, ζ, are the *random* outcomes of $\mathscr{H}$. If to every ζ we assign a real number $X(\zeta)$, we establish a correspondence rule between ζ and R^1, the real line. Such a rule, subject to certain constraints, is called a *random variable*, abbreviated as RV. Thus, a random variable $X(\cdot)$ or simply X is not really a variable but a function whose domain is Ω and whose range is some subset of the real line. Being a function, X generates for every ζ a specific $X(\zeta)$ although for a particular $X(\zeta)$ there may be more than one outcome ζ that produced it. Now consider an event $E_B \subset \Omega (E_B \in \mathscr{F})$.

Through the mapping X, such an event maps into points on the real line (Figure 2.2-1). In particular, the event $\{\zeta\colon X(\zeta) \leq x\}$, often abbreviated $\{X \leq x\}$, will denote an event of unique importance, and we should like to assign a probability to it. As a function of the real variable x, the probability $P[X \leq x] \triangleq F_X(x)$ is called the *cumulative distribution function* (CDF) of X. It is shown in more advanced books [2-1] and [2-2] that in order for $F_X(x)$

†See Appendix D for a brief discussion on measure.

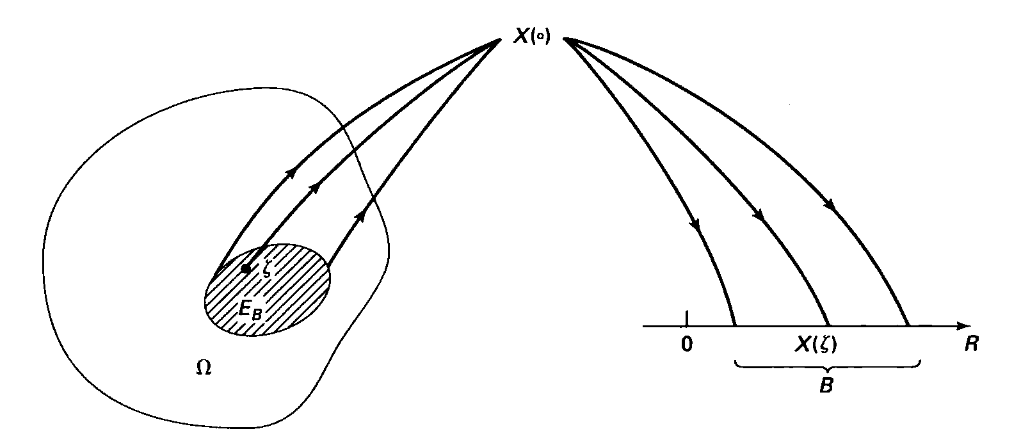


Figure 2.2-1 Symbolic representation of the action of the random variable X.

to be consistent with the axiomatic definition of probability, the function X must satisfy the following: For every Borel set of numbers B, the set $\{\zeta\colon X(\zeta) \in B\}$ must correspond to an event $E_B \in \mathscr{F}$; that is, it must be in the domain of the probability measure P. Stated somewhat more mathematically, this requirement demands that X can be a random variable only if the *inverse image* under X of every Borel subsets in R^1, making up the field $\mathscr{B}$[†] are events. What is an inverse image? Consider an arbitrary Borel set of real numbers B; the set of points E_B in Ω for which $X(\zeta)$ assumes values in B is called the inverse image of the set B under the mapping X. Finally, all sets of engineering interest can be written as countable unions or intersections of events of the form $(-\infty, x]$. The event $\{\zeta\colon X(\zeta) \leq x\} \in \mathscr{F}$ gets mapped under X into $(-\infty, x] \in \mathscr{B}$. *Thus, if X is a random variable, the set of points $(-\infty, x]$ is an event.*

In many if not most scientific and engineering applications, we are not interested in the actual form of X or the specification of the set Ω. For example, we might conceive of an underlying experiment that consists of heating a resistor and observing the positions and velocities of the electrons in the resistor. The set Ω is then the totality of positions and velocities of all N electrons present in the resistor. Let X be the thermal noise current produced by the resistor; clearly $X\colon \Omega \to R^1$ although the form of X, that is, the exceedingly complicated equations of quantum electrodynamics that map from electron positions and velocity configurations to current, is not specified. What we are really interested in is the behavior of X. Thus, although an underlying experiment with sample space Ω may be implied, it is the real line R^1 and its subsets that will hold our interest and figure in our computations. Under the mapping X we have, in effect, generated a new probability space $(R^1, \mathscr{B}, P_X)$, where R^1 is the real line, $\mathscr{B}$ is the Borel σ-field of all subsets of R^1 generated

[†]The σ-field of events defined on Ω is denoted by $\mathscr{F}$. The family of Borel subsets of points on R^1 is denoted by $\mathscr{B}$. For definitions, see Section 1.4 in Chapter 1.

by all the unions, intersections, and complements of the semi-infinite intervals $(-\infty, x]$, and P_X is a set function assigning a number $P_X[A] \geq 0$ to each set $A \in \mathscr{B}$.[†]

In order to assign certain desirable continuity properties to the function $F_X(x)$ at $x = \pm\infty$, we require that the events $\{X = \infty\}$ and $\{X = -\infty\}$ have probability zero. With the latter our specification of a random variable is complete, and we can summarize much of the above discussion in the following definition.

Definition 2.2-1 Let $\mathscr{H}$ be an experiment with sample space Ω. Then the real random variable X is a function whose domain is Ω that satisfies the following: (i) For every Borel set of numbers B, the set $E_B \triangleq \{\zeta \in \Omega, X(\zeta) \in B\}$ is an event and (ii) $P[X = -\infty] = P[X = +\infty] = 0$.

Loosely speaking, when the range of X consists of a countable set of points, X is said to be a discrete random variable; and if the range of X is a continuum, X is said to be continuous. This is a somewhat inadequate definition of discrete and continuous random variables for the simple reason that we often like to take for the range of X the whole real line R^1. Points in R^1 not actually reached by the transformation X with a nonzero probability are then associated with the impossible event.[‡] ■

Example 2.2-1

(*random person*) A person, chosen at random off the street, is asked if he or she has a younger brother. If the answer is *no*, the data is encoded by random variable X as *zero*; if the answer is *yes*, the data is encoded as *one*. The underlying experiment has sample space $\Omega = \{\text{no, yes}\}$, sigma field $\mathscr{F} = [\phi, \Omega, \{\text{no}\}, \{\text{yes}\}]$, and probabilities $P[\phi] = 0$, $P[\Omega] = 1$, $P[\text{no}] = \frac{3}{4}$ (an assumption), $P[\text{yes}] = \frac{1}{4}$. The associated probabilities for X are $P[\phi] = 0$, $P[X \leq \infty] = P[\Omega] = 1$, $P[X = 0] = P[\text{no}] = \frac{3}{4}$, $P[X = 1] = P[\text{yes}] = \frac{1}{4}$. Take any x_1, x_2 and consider, for example, the probabilities that X lies in sets of the type $[x_1, x_2]$, $[x_1, x_2)$, or $(x_1, x_2]$. Thus,

$$P[3 \leq X \leq 4] = P[\phi] = 0$$

$$P[0 \leq X < 1] = P[\text{no}] = \tfrac{3}{4}$$

$$P[0 \leq X \leq 2] = P[\Omega] = 1$$

$$P[0 < X \leq 1] = P[\text{yes}] = \tfrac{1}{4},$$

and so on. Thus, every set $\{X = x\}$, $\{x_1 \leq X < x_2\}$, $\{X \leq x_2\}$, and so forth is related to an event defined on Ω. Hence X is a random variable.

[†]The extraordinary advantage of dealing with random variables is that a single pointwise function, that is, the cumulative distribution function $F_X(x)$, can replace the set function $P_X[\cdot]$ that may be extremely cumbersome to specify, since it must be specified for every event (set) $A \in \mathscr{B}$. See Section 2.3.

[‡]An alternative definition is the following: X is discrete if $F_X(x)$ is a staircase-type function, and X is continuous if $F_X(x)$ is a continuous function. Some random variables cannot be classified as discrete or continuous; they are discussed in Section 2.5.

Example 2.2-2

(*random bus arrival time*) A bus arrives at random in $[0, T]$; let t denote the time of arrival. The sample space Ω is $\Omega = \{t: t \in [0, T]\}$. A random variable X is defined by

$$X(t) = \begin{cases} 1, & t \in \left[\frac{T}{4}, \frac{T}{2}\right], \\ 0, & \text{otherwise.} \end{cases}$$

Assume that the arrival time is uniform over $[0, T]$. We can now ask and compute what is $P[X(t) = 1]$ or $P[X(t) = 0]$ or $P[X(t) \leq 5]$.

Example 2.2-3

(*drawing from urn*) An urn contains three colored balls. The balls are colored white (W), black (B), and red (R), respectively. The experiment consists of choosing a ball at random from the urn. The sample space is $\Omega = \{W, B, R\}$. The random variable X is defined by

$$X(\zeta) = \begin{cases} \pi, & \zeta = W \text{ or } B, \\ 0, & \zeta = R. \end{cases}$$

We can ask and compute the probability $P[X \leq x_1]$, where x_1 is any number. Thus, $\{X \leq 0\} = \{R\}$, $\{2 \leq X < 4\} = \{W, B\}$. The computation of the associated probabilities is left as an exercise.

Example 2.2-4

(*wheel of chance*) A spinning wheel and pointer has 50 sectors numbered $n = 0, 1, \ldots, 49$. The experiment consists of spinning the wheel. Because the players are interested only in *even* or *odd* outcomes, they choose $\Omega = \{\textit{even}, \textit{odd}\}$ and the only events in the σ-field are $\{\phi, \Omega, \textit{even}, \textit{odd}\}$. Let $X = n$, that is, if n shows up, X assumes that value. Is X a random variable? Note that the inverse image of the set $\{2, 3\}$ is not an event. Hence X is not a valid random variable on this probability space because it is not a function on Ω.

2.3 CUMULATIVE DISTRIBUTION FUNCTION

In Example 2.2-1 the induced event space under X includes $\{0, 1\}$, $\{0\}$, $\{1\}$, ϕ, for which the probabilities are $P[X = 0 \text{ or } 1] = 1$, $P[X = 0] = \frac{3}{4}$, $P[X = 1] = \frac{1}{4}$, and $P[\phi] = 0$. From these probabilities, we can infer any other probabilities such as $P[X \leq 0.5]$. In many cases it is awkward to write down $P[\cdot]$ for every event. For this reason we introduce a pointwise probability function called the *cumulative distribution function* CDF. The CDF is a function of x, which contains all the information necessary to compute $P[E]$ for any E in the Borel field of events. The CDF, $F_X(x)$, is defined by

$$F_X(x) = P[\{\zeta: X(\zeta) \leq x\}] = P_X[(-\infty, x]]. \tag{2.3-1}$$

Equation 2.3-1 is read as "the set of all outcomes ζ in the underlying sample space such that the function $X(\zeta)$ assumes values less than or equal to x." Thus, there is a subset of outcomes $\{\zeta\colon X(\zeta) \leq x\} \subset \Omega$ that the mapping $X(\cdot)$ generates as the set $[-\infty, x] \subset R^1$. The sets $\{\zeta\colon X(\zeta) \leq x\} \subset \Omega$ and $[-\infty, x] \subset R^1$ are equivalent events. We shall frequently leave out the dependence on the underlying sample space and write merely $P[X \leq x]$ or $P[a < X \leq b]$.

For the present we shall denote random variables by capital letters, that is, X, Y, Z, and the values they can take by lowercase letters x, y, z. The subscript X on $F_X(x)$ associates it with the random variable for which it is the CDF. Thus, $F_X(y)$ means the CDF of random variable X evaluated at the real number y and thus equals the probability $P[X \leq y]$. If $F_X(x)$ is discontinuous at a point, say, x_o, then $F_X(x_o)$ will be taken to mean the value of the CDF immediately to the right of x_o(we call the *continuity from the right*).

Properties† of $F_X(x)$

(i) $F_X(\infty) = 1$, $F_X(-\infty) = 0$.
(ii) $x_1 \leq x_2 \rightarrow F_X(x_1) \leq F_X(x_2)$, that is, $F_X(x)$ is a nondecreasing function of x.
(iii) $F_X(x)$ is continuous from the right, that is,

$$F_X(x) = \lim_{\varepsilon \to 0} F_X(x + \varepsilon) \qquad \varepsilon > 0.$$

Proof of (ii) Consider the event $\{x_1 < X \leq x_2\}$ with $x_2 > x_1$. The set $[x_1, x_2]$ is nonempty and $\in \mathscr{B}$. Hence

$$0 \leq P[x_1 < X \leq x_2] \leq 1.$$

But

$$\{X \leq x_2\} = \{X \leq x_1\} \cup \{x_1 < X \leq x_2\}$$

and

$$\{X \leq x_1\} \cap \{x_1 < X \leq x_2\} = \phi.$$

Hence

$$F_X(x_2) = F_X(x_1) + P[x_1 < X \leq x_2]$$

or

$$P[x_1 < X \leq x_2] = F_X(x_2) - F_X(x_1) \geq 0 \text{ for } x_2 > x_1. \tag{2.3-2}$$

We leave it to the reader to establish the following results:

$$P[a \leq X \leq b] = F_X(b) - F_X(a) + P[X = a];$$
$$P[a < X < b] = F_X(b) - P[X = b] - F_X(a);$$
$$P[a \leq X < b] = F_X(b) - P[X = b] - F_X(a) + P[x = a].$$

†Properties (i) and (iii) require proof. This is furnished with the help of extended axioms in Chapter 8. Also see Davenport [2-3, Chapter 4].

Example 2.3-1

(*parity bits*) The experiment consists of observing the voltage X of the parity bit in a word in computer memory. If the bit is *on*, then $X = 1$; if *off* then $X = 0$. Assume that the *off* state has probability q and the *on* state has probability $1 - q$. The sample space has only two points: $\Omega = \{\text{off, on}\}$.

Computation of $F_X(x)$

(i) $x < 0$: The event $\{X \leq x\} = \phi$ and $F_X(x) = 0$.

(ii) $0 \leq x < 1$: The event $\{X \leq x\}$ is equivalent to the event $\{\text{off}\}$ and excludes the event $\{\text{on}\}$.

$$X(\text{on}) = 1 > x$$

$$X(\text{off}) = 0 \leq x.$$

Hence $F_X(x) = q$.

(iii) $x \geq 1$: The event $\{X \leq x\} =$ is the certain event since

$$X(\text{on}) = 1 \leq x$$

$$X(\text{off}) = 0 \leq x.$$

The solution is shown in Figure 2.3-1.

Example 2.3-2

(*waiting for a bus*) A bus arrives at random in $(0, T]$. Let the random variable X denote the time of arrival. Then clearly $F_X(t) = 0$ for $t \leq 0$ and $F_X(T) = 1$ because the former is the probability of the impossible event while the latter is the probability of the certain event. Suppose it is known that the bus is equally likely or *uniformly* likely to come at any time within $(0, T]$. Then

$$F_X(t) = \begin{cases} 0, & t \leq 0, \\ \dfrac{t}{T}, & 0 < t \leq T, \\ 1, & t > T. \end{cases} \tag{2.3-3}$$

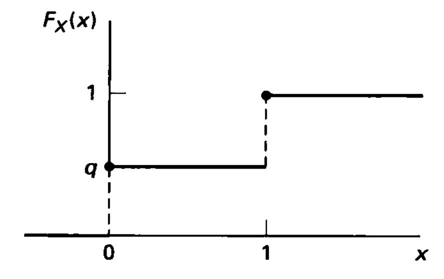


Figure 2.3-1 Cumulative distribution function associated with the parity bit observation experiment.

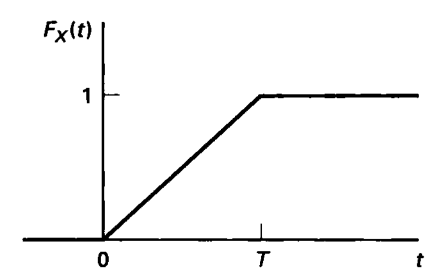


Figure 2.3-2 Cumulative distribution function of the uniform random variable X of Example 2.3-2.

Actually Equation 2.3-3 defines "equally likely," not the other way around. The CDF is shown in Figure 2.3-2. In this case we say that X is *uniformly* distributed.

If $F_X(x)$ is a continuous function of x, then

$$F_X(x) = F_X(x^-). \tag{2.3-4}$$

However, if $F_X(x)$ is discontinuous at the point x, then, from Equation 2.3-2,

$$\begin{aligned} F_X(x) - F_X(x^-) &= P[x^- < X \leq x] \\ &= \lim_{\varepsilon \to 0} P[x - \varepsilon < X \leq x] \\ &\triangleq P[X = x]. \end{aligned} \tag{2.3-5}$$

Typically $P[X = x]$ is a discontinuous function of x; it is zero whenever $F_X(x)$ is continuous and nonzero only at discontinuities in $F_X(x)$.

Example 2.3-3

(*binomial distribution function*) Compute the CDF for a binomial random variable X with parameters (n, p).

Solution Since X takes on only discrete values, that is, $X \in \{0, 1, 2, \ldots, n\}$, the event $\{X \leq x\}$ is the same as $\{X \leq [x]\}$, where $[x]$ is the largest integer equal to or smaller than x. Then $F_X(x)$ is given by the stepwise constant function

$$F_X(x) = \sum_{j=0}^{[x]} \binom{n}{k} p^j (1-p)^{n-j}.$$

For $p = 0.6$, $n = 4$, the CDF has the appearance of a staircase function as shown in Figure 2.3-3.

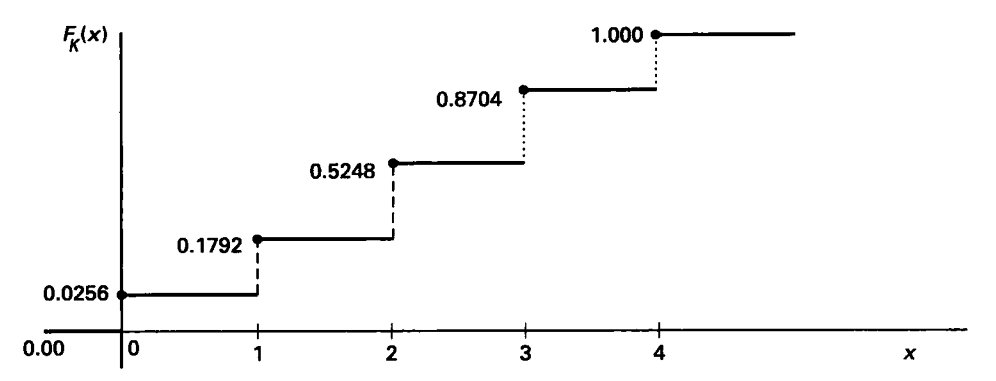


Figure 2.3-3 Cumulative distribution function for a binomial RV with $n = 4$, $p = 0.6$.

Example 2.3-4

(*computing binomial probabilities*) Using the results of Example 2.3-3, compute the following:

(a) $P[1.5 < X < 3]$;
(b) $P[0 \leq X \leq 3]$;
(c) $P[1.2 < X \leq 1.8]$;
(d) $P[1.99 \leq < 3]$.

Solution

(a) $P[1.5 < X < 3] = F_X(3) - P[X = 3] - F_X(1.5)$
$= 0.8704 - 0.3456 - 0.1792 = 0.3456$;

(b) $P[0 \leq X \leq 3] = F_X(3) - F_X(0) + P[X = 0]$
$= 0.8704 - 0.0256 + 0.0256 = 0.8704$;

(c) $P[1.2 < X \leq 1.8] = F_X(1.8) - F_X(1.2)$
$= 0.1792 - 0.1792 = 0$;

(d) $P[1.99 \leq X < 3] = F_X(3) - P[X = 3] - F_X(1.99) + P[X = 1.99]$
$= 0.8704 - 0.3456 - 0.1792 + 0 = 0.3456$

Note that even for a discrete RV, we have taken the CDF to be a function of a continuous variable, x in this example. However, for a discrete RV, it is sometimes simpler (but more restrictive) to consider the CDF to be discrete also. Let X be a discrete RV taking on values $\{x_k\}$ with probability mass function (PMF) $P_X(x_k)$. Then the *discrete* CDF would only be defined on the values $\{x_k\}$ also. Assuming that these values are an increasing set, that is, $x_k < x_{k+1}$ for all k, the discrete CDF would be

$$F_X(x_k) \triangleq \sum_{j=-\infty}^{k} P(x_j) \quad \text{for all } k.$$

In this format, we compute the CDF only at points corresponding to the countable outcomes of the sample space.

Looking again at the binomial example $b(n,p)$ above, but using the discrete CDF, we would say the RV K takes on values in the set $\{0 \leq k \leq 4\}$ with the discrete CDF

$$F_K(k) = \sum_{j=0}^{k} (0.6)^j (1-0.6)^{n-j} \quad \text{for} \quad 0 \leq k \leq 4.$$

While this is more natural for a discrete RV, the reader will note that the discrete CDF cannot be used to evaluate probabilities such as $P[1.5 < K < 3]$ since it cannot be evaluated at 1.5. For this reason, we generally will consider CDFs as defined for a continuous domain, even though the RV in question might be discrete valued.

2.4 PROBABILITY DENSITY FUNCTION (pdf)

If $F_X(x)$ is continuous and differentiable, the pdf is computed from

$$f_X(x) = \frac{dF_X(x)}{dx}. \tag{2.4-1}$$

Properties. If $f_X(x)$ exists, then

(i) $f_X(x) \geq 0.$ (2.4-2)

(ii) $\int_{-\infty}^{\infty} f_X(\xi)d\xi = F_X(\infty) - F_X(-\infty) = 1.$ (2.4-3)

(iii) $F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi = P[X \leq x].$ (2.4-4)

(iv) $F_X(x_2) - F_X(x_1) = \int_{-\infty}^{x_2} f_X(\xi)d\xi - \int_{-\infty}^{x_1} f_X(\xi)d\xi$

$$= \int_{x_1}^{x_2} f_X(\xi)d\xi = P[x_1 < X \leq x_2]. \tag{2.4-5}$$

Interpretation of $f_X(x)$.

$$P[x < X \leq x + \Delta x] = F_X(x + \Delta x) - F_X(x).$$

If $F_X(x)$ is continuous in its first derivative then, for sufficiently small Δx,

$$F_X(x + \Delta x) - F_X(x) = \int_{x}^{x+\Delta x} f(\xi)d\xi \simeq f_X(x)\Delta x.$$

Hence for small Δx

$$P[x < X \leq x + \Delta x] \simeq f_X(x)\Delta x. \tag{2.4-6}$$

Observe that if $f_X(x)$ exists, meaning that it is bounded and has at most a finite number of discontinuities then $F_X(x)$ is continuous and therefore, from Equation 2.3-5, $P[X = x] = 0$.

The univariate Normal (Gaussian[†]) pdf. The pdf is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}, -\infty < x < +\infty. \tag{2.4-7}$$

There are two distinct parameters: the *mean* μ and the *standard deviation* $\sigma(> 0)$. (Note that σ^2 is called the *variance*). We show that this density is valid by integrating over all x as follows

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy, \quad \text{with the substitution } y \triangleq \frac{x-\mu}{\sigma},$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{+\infty} e^{-\frac{y^2}{2}} dy = \frac{2}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}} = \frac{2\sqrt{\pi}}{2\sqrt{\pi}} = 1,$$

where we make use of the known integral

$$\int_0^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{\frac{\pi}{2}}.$$

Now the Gaussian (Normal) random variable is very common in applications and a special notation is used to specify it. We often say that X is distributed as $N(\mu, \sigma^2)$ or write $X : N(\mu, \sigma^2)$ to specify this distribution.[‡]

For any random variable with a well-defined pdf, we can in general compute the mean and variance (the square of the standard deviation), if it exists, from the two formulas

$$\mu \triangleq \int_{-\infty}^{\infty} x f_X(x) dx \tag{2.4-8}$$

and

$$\sigma^2 \triangleq \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx. \tag{2.4-9}$$

We will defer to Chapter 4 the proof that the parameters we call μ and σ^2 in the Gaussian distribution are actually the true mean and variance as defined generally in these two equations.

For discrete random variables, we compute the mean and variance from the sums

$$\mu \triangleq \sum_{i=-\infty}^{\infty} x_i P_X(x_i) \tag{2.4-10}$$

[†]After the German mathematician/physicist Carl F. Gauss (1777–1855).

[‡]The reader may note that capital letter on the word *Normal*. We use this choice to make the reader aware that while Gaussian or Normal is very common, it is not *normal* or ubiquitous in the everyday sense.

and

$$\sigma^2 \triangleq \sum_{i=-\infty}^{\infty} (x_i - \mu)^2 P_X(x_i). \tag{2.4-11}$$

Here are some simple examples of the computation of mean and variance.

Example 2.4-1

Let $f_X(x) = 1$, for $0 < x \leq 1$ and zero elsewhere. This pdf is a special case of the *uniform law* discussed below. The mean is computed as

$$\mu = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x\, dx = 0.5$$

and the variance is computed as

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_X dx = \int_0^1 (x - 0.5)^2 dx = 1/12.$$

Example 2.4-2

Suppose we are given that $P_X(0) = P_X(2) = 0.25$, $P_X(1) = 0.5$, and zero elsewhere. For this discrete RV, we use Equations 2.4-10 and 2.4-11 to obtain

$$\mu = 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1$$

and

$$\sigma^2 = (0-1)^2 \times 0.25 + (1-1)^2 \times 0.5 + (2-1)^2 \times 0.25 = 0.5.$$

The mean and variance are common examples of statistical moments, whose discussion is postponed till Chapter 4. The Normal pdf is shown in Figure 2.4-1.

The Normal pdf is widely encountered in all branches of science and engineering as well as in social and demographic studies. For example, the IQ of children, the heights of men (or women), and the noise voltage produced by a thermally agitated resistor are all postulated to be approximately Normal over a large range of values.

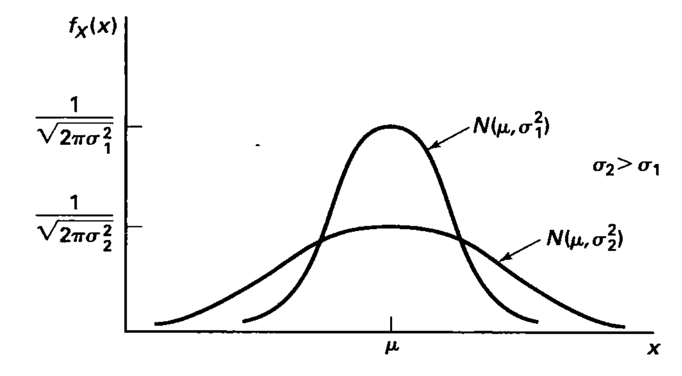


Figure 2.4-1 The Normal pdf.

Conversion of the Gaussian pdf to the standard Normal. Suppose we are given X: $N(\mu, \sigma^2)$ and must evaluate $P[a < X \leq b]$. We have

$$P[a < X \leq b] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b e^{-\frac{1}{2}\left[\frac{x-u}{\sigma}\right]^2} dx.$$

With $\beta \triangleq (x - \mu)/\sigma$, $d\beta = (1/\sigma)dx$, $b' \triangleq (b - \mu)/\sigma$, $a' \triangleq (a - \mu)/\sigma$, we obtain

$$\begin{aligned} P[a < X \leq b] &= \frac{1}{\sqrt{2\pi}} \int_{a'}^{b'} e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_0^{b'} e^{-\frac{1}{2}x^2} dx - \frac{1}{\sqrt{2\pi}} \int_0^{a'} e^{-\frac{1}{2}x^2} dx. \end{aligned}$$

The function

$$\text{erf}(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}t^2} dt \tag{2.4-12}$$

is sometimes called the error function [erf(x)] although other definitions of erf(x) exist.† The erf(x) is tabulated in Table 2.4-1 and is plotted in Figure 2.4-2.

Hence if X:$N(\mu, \sigma^2)$, then

$$P[a < X \leq b] = \text{erf}\left(\frac{b - \mu}{\sigma}\right) - \text{erf}\left(\frac{a - \mu}{\sigma}\right). \tag{2.4-13}$$

Example 2.4-3

(*resistor tolerance*) Suppose we choose a resistor with resistance R from a batch of resistors with parameters $\mu = 1000$ ohms and $\sigma = 200$ ohms. What is the probability that R will have a value between 900 and 1100 ohms?

Solution Assuming that R: $N[1000, (200)^2]$ we compute from Equation 2.4-13

$$P[900 < R \leq 1100] = \text{erf}(0.5) - \text{erf}(-0.5).$$

But $\text{erf}(-x) = -\text{erf}(x)$ (deduced from Equation 2.4-12). Hence

$$P[900 < R \leq 1100] = 0.38.$$

---

†For example, a widely used definition of erf(x) is $\text{erf}_2(x) \triangleq (2/\sqrt{\pi}) \int_0^x e^{-t^2} dt$, which is used in MATLAB. The relation between these two erf's is $\text{erf}(x) = \frac{1}{2}\text{erf}_2(x/\sqrt{2})$.

Table 2.4-1 Selected Values of erf(x)

$$\operatorname{erf}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\left(-\frac{1}{2}t^2\right) dt$$

x	erf(x)	x	erf(x)
0.05	0.01994	2.05	0.47981
0.10	0.03983	2.10	0.48213
0.15	0.05962	2.15	0.48421
0.20	0.07926	2.20	0.48609
0.25	0.09871	2.25	0.48777
0.30	0.11791	2.30	0.48927
0.35	0.13683	2.35	0.49060
0.40	0.15542	2.40	0.49179
0.45	0.17364	2.45	0.49285
0.50	0.19146	2.50	0.49378
0.55	0.20884	2.55	0.49460
0.60	0.22575	2.60	0.49533
0.65	0.24215	2.65	0.49596
0.70	0.25803	2.70	0.49652
0.75	0.27337	2.75	0.49701
0.80	0.28814	2.80	0.49743
0.85	0.30233	2.85	0.49780
0.90	0.31594	2.90	0.49812
0.95	0.32894	2.95	0.49840
1.00	0.34134	3.00	0.49864
1.05	0.35314	3.05	0.49884
1.10	0.36433	3.10	0.49902
1.15	0.37492	3.15	0.49917
1.20	0.38492	3.20	0.49930
1.25	0.39434	3.25	0.49941
1.30	0.40319	3.30	0.49951
1.35	0.41149	3.35	0.49958
1.40	0.41924	3.40	0.49965
1.45	0.42646	3.45	0.49971
1.50	0.43319	3.50	0.49976
1.55	0.43942	3.55	0.49980
1.60	0.44519	3.60	0.49983
1.65	0.45052	3.65	0.49986
1.70	0.45543	3.70	0.49988
1.75	0.45993	3.75	0.49990
1.80	0.46406	3.80	0.49992
1.85	0.46783	3.85	0.49993
1.90	0.47127	3.90	0.49994
1.95	0.47440	3.95	0.49995
2.00	0.47724	4.00	0.49996

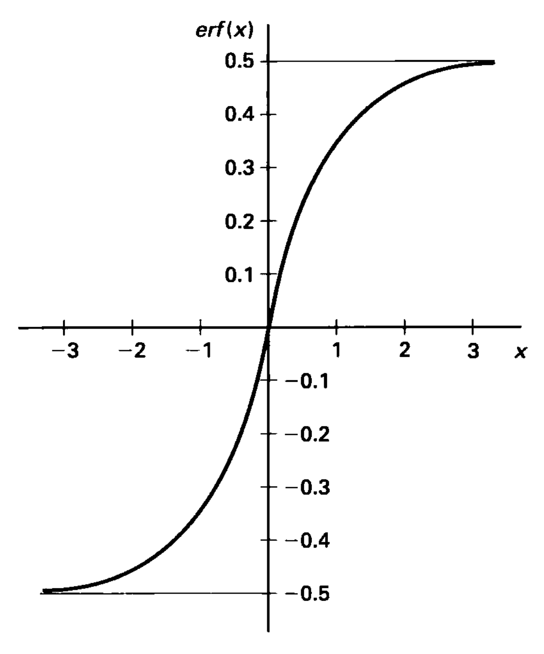


Figure 2.4-2 erf(x) versus x.

Using Figure 2.4-3 as an aid in our reasoning, we readily deduce the following for $X: N(0,1)$. Assume $x > 0$; then

$$P[X \leq x] = \frac{1}{2} + \text{erf}(x), \tag{2.4-14a}$$

$$P[X > -x] = \frac{1}{2} + \text{erf}(x), \tag{2.4-14b}$$

$$P[X > x] = \frac{1}{2} - \text{erf}(x), \tag{2.4-14c}$$

$$P[-x < X \leq x] = 2\,\text{erf}(x), \tag{2.4-14d}$$

$$P[|X| > x] = 1 - 2\,\text{erf}(x). \tag{2.4-14e}$$

Example 2.4-4

(*manufacturing*) A metal rod is nominally 1 meter long, but due to manufacturing imperfections, the actual length L is a Gaussian random variable with mean $\mu = 1$ and standard deviation $\sigma = 0.005$. What is the probability that the rod length L lies in the interval $[0.99, 1.01]$? Since the random variable $L{:}N(1, (0.005)^2)$, we have

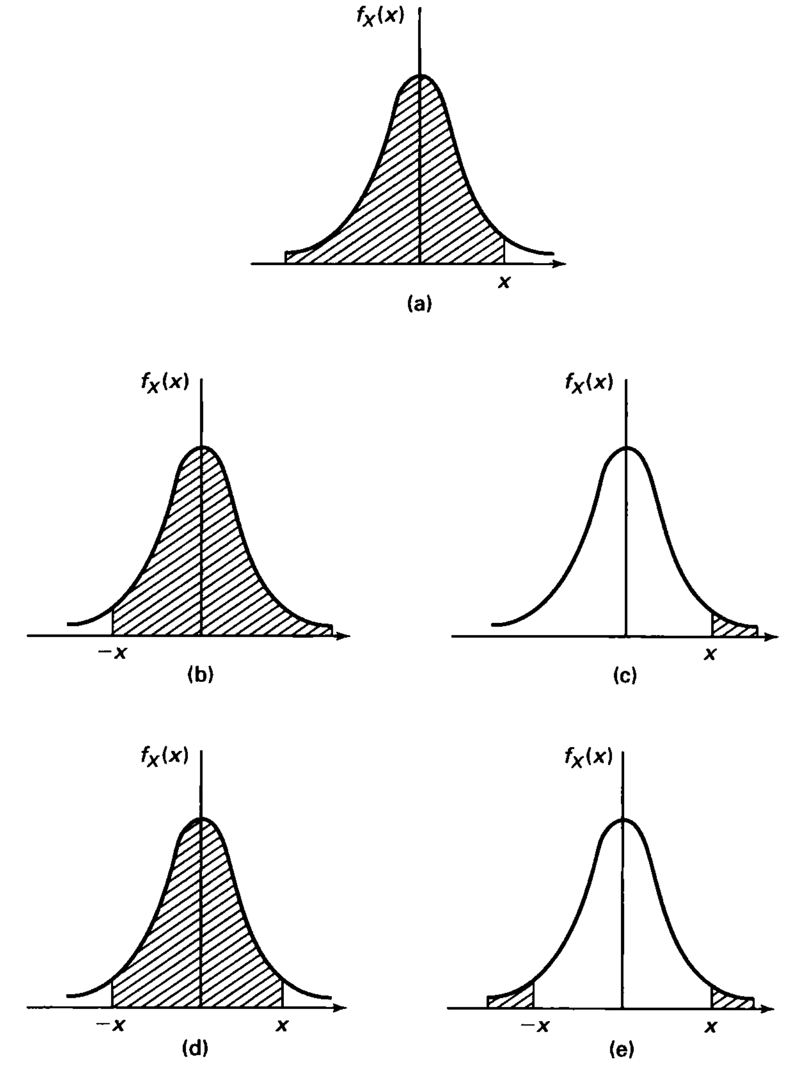


Figure 2.4-3 The areas of the shaded region under curves are (a) $P[X \leq x]$; (b) $P[X > -x]$; (c) $P[X > x]$; (d) $P[-x < X \leq x]$; and (e) $P[|X| > x]$.

$$
\begin{aligned}
P[0.99 < L \leq 1.01] &= \int_{0.99}^{1.01} \frac{1}{\sqrt{2\pi}(0.005)} e^{-\frac{1}{2}\left(\frac{x-1.00}{0.005}\right)^2} dx \\
&= \int_{\frac{0.99-1.00}{0.005}}^{\frac{1.01-1.00}{0.005}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx
\end{aligned}
$$

$$= \int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

$$= 2\,\text{erf}(2) \doteq 2 \times 0.4772 \qquad \text{(from Table 2.4-1)}$$

$$= 0.954.$$

Four Other Common Density Functions

1. *Rayleigh* ($\sigma > 0$):

$$f_X(x) = \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} u(x), \tag{2.4-15}$$

where the continuous unit-step function is defined as

$$u(x) \triangleq \begin{cases} 1, & 0 \le x < \infty, \\ 0, & -\infty < x < 0. \end{cases}$$

Thus, $f_X(x) = 0$ for $x < 0$. Examples of where the Rayleigh pdf shows up are in rocket-landing errors, random fluctuations in the envelope of certain waveforms, and radial distribution of misses around the bull's-eye at a rifle range.

2. *Exponential* ($\mu > 0$):

$$f_X(x) = \frac{1}{\mu} e^{-x/\mu} u(x). \tag{2.4-16}$$

The exponential law occurs, for example, in waiting-time problems, in calculating lifetime of machinery, and in describing the intensity variations of incoherent light.

3. *Uniform* ($b > a$):

$$f_X(x) = \frac{1}{b-a} \qquad a < x < b$$
$$= 0 \qquad \text{otherwise.} \tag{2.4-17}$$

The uniform pdf is used in communication theory, in queueing models, and in situations where we have no *a priori* knowledge favoring the distribution of outcomes except for the end points; that is, we don't know when a business call will come but it must come, say, between 9 A.M. and 5 P.M. We sometimes use the notation $U(a, b)$ to denote a uniform distribution lower-bounded by a and upper-bound by b.

The three pdf's are shown in Figure 2.4-4.

4. *Laplacian*: The pdf is defined by

$$f_X(x) = \frac{c}{2} e^{-c|x|}, \quad -\infty < x < \infty \quad c > 0. \tag{2.4-18}$$

The Laplacian is widely used in speech and image processing to model *adjacent-sample difference* and is the difference in signal level from a sample point and its neighbor. Since

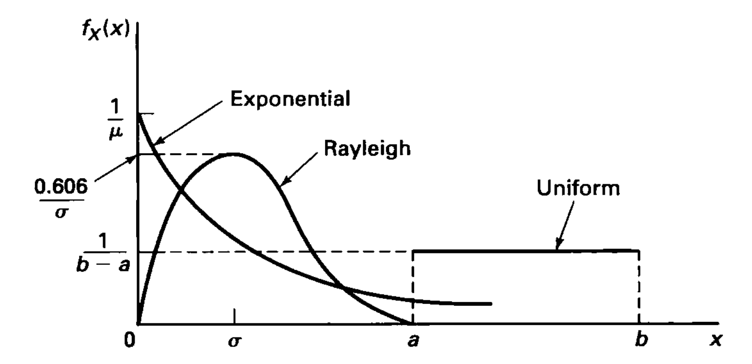


Figure 2.4-4 The Rayleigh, exponential, and uniform pdf's.

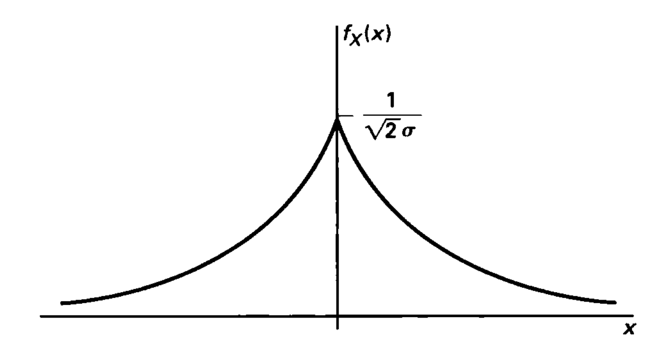


Figure 2.4-5 The Laplacian pdf used in computer analysis of speech and images.

the levels of the sample point and its neighbor are often the same, the Laplacian peaks at zero. The Laplacian pdf is sometime written as

$$f_X(x) = \frac{1}{\sqrt{2}\sigma} \exp[-\sqrt{2}|x|/\sigma], \quad -\infty < x < \infty \quad \sigma > 0, \tag{2.4-19}$$

where σ is the standard deviation of the Laplacian RV X. Precisely what this means will be explained in Chapter 4. The Laplacian pdf is shown in Figure 2.4-5. In image compression, the Laplacian model is appropriate for the so-called "AC coefficients" that arise after a decorrelating transform called the DCT† which is applied on 8×8 blocks of pixels.

Example 2.4-5

(*radiated power*) The received power W on a cell phone at a certain distance from the base station is found to follow a Rayleigh distribution with parameter $\sigma = 1$ milliwatt. What

†DCT stands for discrete cosine transform and is a variation on the DFT used in signal analysis. A 2-D version is used for images, consisting of a 1-D DCT on the rows followed by a 1-D transform on the columns.

is the probability that the power W is less than 0.8 milliwatts? Since the power can be modeled by the Rayleigh random variable, we have

$$
\begin{aligned}
P[W \leq 0.8] &= \int_{0.0}^{0.8} x e^{-\frac{x^2}{2}} dx, \quad \text{since } \sigma^2 = 1, \\
&= \int_{0.0}^{0.8} e^{-\frac{x^2}{2}} d(\frac{1}{2}x^2) \\
&= \int_{0.0}^{0.32} e^{-y} dy, \quad \text{with the substitution } y \triangleq \frac{1}{2}x^2, \\
&= 1 - e^{-0.32} \simeq 0.29.
\end{aligned}
$$

Example 2.4-6

(*image compression*) In designing the quantizer for a JPEG image compression system, we need to know what the range should be for the transformed AC coefficients. Using the Laplacian model with parameter σ for such a coefficient X, what is the probability of the event $\{|X| > k\sigma\}$ as a function of $k = 1, 2, 3, ...$? If we then make this probability sufficiently low, by choice of k, we will design the quantizer for the range $[-k\sigma, +k\sigma]$ and only saturate the quantizer occasionally. We need to calculate

$$
\begin{aligned}
P[|X| > k\sigma] &= \int_{k\sigma}^{\infty} \frac{1}{\sqrt{2}\sigma} \exp\left(-\sqrt{2}x/\sigma\right) dx + \int_{-\infty}^{-k\sigma} \frac{1}{\sqrt{2}\sigma} \exp\left(+\sqrt{2}x/\sigma\right) dx \\
&= 2 \int_{k\sigma}^{\infty} \frac{1}{\sqrt{2}\sigma} \exp\left(-\sqrt{2}x/\sigma\right) dx \\
&= 2 \int_{k}^{\infty} \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}y\right) dy, \quad \text{with } y \triangleq x/\sigma, \\
&= 2(\frac{1}{2} \exp\left(-\sqrt{2}y\right) \Big|_{k}^{\infty} \\
&= \exp\left(-\sqrt{2}k\right).
\end{aligned}
$$

For $k = 2$, we get probability 0.059 and for $k = 5$ we get 0.85×10^{-3}, or about one in a thousand coefficients.

Table 2.4-2 lists some common continuous random variables with their probability densities and distribution functions.

More Advanced Density Functions

5. Chi-square (n an integer)

$$
f_X(x) = K_\chi x^{(\frac{n}{2})-1} e^{-\frac{x}{2}} u(x), \tag{2.4-20}
$$

Table 2.4-2 Common Continuous Probability Densities and Distribution Functions

Family	pdf $f_X(x)$	CDF $F_X(x)$
Uniform $U(a,b)$	$\frac{1}{b-a}\,[u(x-a)-u(x-b)]$	$\begin{cases} 0, & x<a, \\ \frac{x-a}{b-a}, & a\le x<b \\ 1, & b\le x \end{cases}$,
Exponential $\mu>0$	$\frac{1}{\mu}e^{-x/\mu}\,u(x)$	$\begin{cases} 0, & x<0, \\ 1-e^{-x/\mu}, & x\ge 0 \end{cases}$
Gaussian $N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2]$	$\frac{1}{2}+\text{erf}(\frac{x-\mu}{\sigma})$
Laplacian $\sigma>0$	$\frac{1}{\sqrt{2}\sigma}\exp[-\sqrt{2}\lvert x\rvert/\sigma]$	$\frac{1}{2}[1+\text{sgn}(x)(1-\exp(-\sqrt{2}\lvert x\rvert/\sigma))]$
Rayleigh $\sigma>0$	$\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}u(x)$	$\left[1-e^{-x^2/2\sigma^2}\right]u(x)$

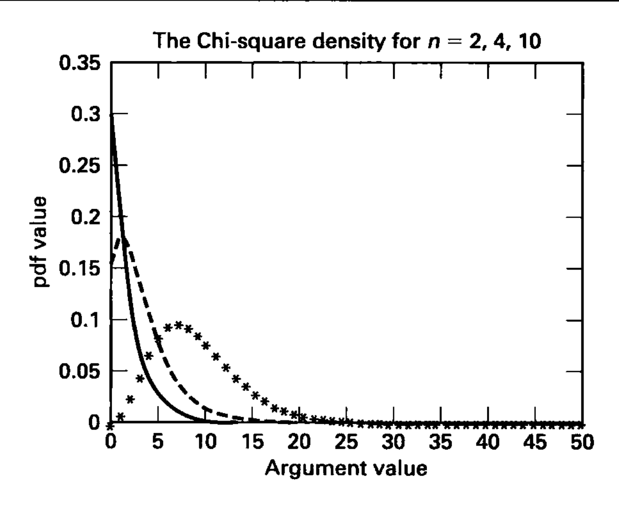


Figure 2.4-6 The Chi-square probability density function for $n = 2$ (solid), $n = 4$ (dashed), and $n = 10$ (stars). Note that for larger values of n, the shape approaches that of a Normal pdf with a positive mean-parameter μ.

where the normalizing constant K_χ is computed as $K_\chi = \frac{1}{2^{n/2}\Gamma(n/2)}$ and $\Gamma(\cdot)$ is the *Gamma function* discussed in Appendix B. The Chi-square pdf is shown in Figure 2.4-6.

6. Gamma: $(b>0, c>0)$

$$f_X(x) = K_\gamma x^{b-1}e^{-cx}u(x), \tag{2.4-21}$$

where $K_\gamma = c^b/\Gamma(b)$.

7. Student-t: (n an integer)

$$f_X(x) = K_{st}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, -\infty < x < \infty \tag{2.4-22}$$

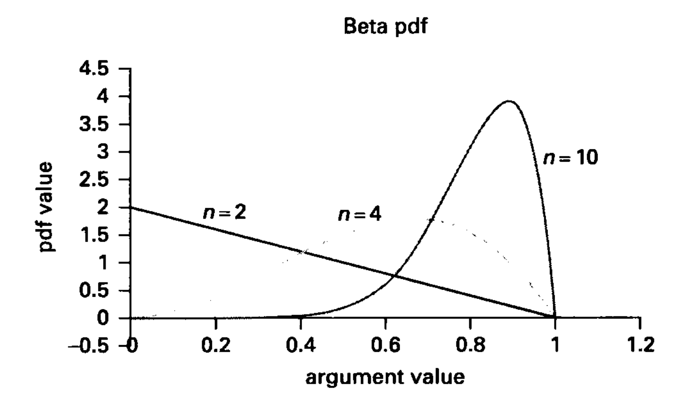


Figure 2.4-7 The beta pdf shown for $\beta = 1, \alpha = n - 2$, and various values of n. When $\beta = 0, \alpha = 0$ the beta pdf becomes uniformly distributed over $0 < x < 1$.

where

$$K_{st} = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{\pi n}}.$$

The Chi-square and Student-t densities are widely used in statistics.† We shall encounter these densities later in the book. The gamma density is mother to other densities. For example with $b = 1$, there results the exponential density; and with $b = n/2$ and $c = 1/2$, there results the Chi-square density.

8. Beta ($\alpha > 0, \beta > 0$):

$$f_X(x; \alpha, \beta) = \left\{ \begin{array}{c} \frac{(\alpha+\beta+1)!}{\alpha!\beta!} x^{\alpha}(1-x)^{\beta}, \ 0 < x < 1, \\ 0 \text{ , else.} \end{array} \right\}$$

The beta distribution is a two-parameter family of functions that appears in statistics. It is shown in Figure 2.4-7.

There are other pdf's of importance in engineering and science, and we shall encounter some of them as we continue our study of probability. They all, however, share the properties that

†The Student-t distribution is so named because its discoverer W. S. Gossett (1876–1937) published his papers under the name "Student." Gossett, E. S. Pearson, R. A. Fisher, and J. Neyman are regarded as the founders of modern statistics.

$$f_X(x) \geq 0 \tag{2.4-23}$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1. \tag{2.4-24}$$

When $F_X(x)$ is not continuous, strictly speaking, its finite derivative does not exist and, therefore, the pdf doesn't exist. The question of what probability function is useful in describing X depends on the classification of X. We consider this next.

2.5 CONTINUOUS, DISCRETE, AND MIXED RANDOM VARIABLES

If $F_X(x)$ is continuous for every x and its derivative exists everywhere except at a countable set of points, then we say that X is a *continuous* RV. At points x where $F'_X(x)$ exists, the pdf is $f_X(x) = F'_X(x)$. At points where $F_X(x)$ is continuous, but $F'_X(x)$ is discontinuous, we can assign any positive number to $f_X(x)$; $f_X(x)$ will then be defined for every x, and we are free to use the following important formulas:

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi, \tag{2.5-1}$$

$$P[x_1 < X \leq x_2] = \int_{x_1}^{x_2} f_X(\xi)d\xi, \tag{2.5-2}$$

and

$$P[B] = \int_{\xi:\ \xi \in B} f_X(\xi)d\xi, \tag{2.5-3}$$

where, in Equation 2.5-3, $B \in \mathscr{B}$, that is, B is an event. Equation 2.5-3 follows from the fact that for a continuous random variable, events can be written as a union of disjoint intervals in R. Thus, for example, let $B = \{\xi\colon \xi \in \cup_{i=1}^{n} I_i, I_i I_j = \phi \text{ for } i \neq j\}$, where $I_i = (a_i, b_i]$. Then clearly,

$$\begin{aligned} P[B] &= \int_{a_1}^{b_1} f_X(\xi)d\xi + \int_{a_2}^{b_2} f_X(\xi)d\xi + \ldots + \int_{a_n}^{b_n} f_X(\xi)d\xi \\ &= \int_{\xi:\ \xi \in B} f_X(\xi)d\xi. \end{aligned} \tag{2.5-4}$$

A *discrete* random variable has a staircase type of distribution function (Figure 2.5-1).

A probability measure for discrete RV is the *probability mass function*† (PMF). The PMF $P_X(x)$ of a (discrete) random variable X is defined as

†Like mass, probability is nonnegative and conserved. Hence the term *mass* in probability *mass* function.

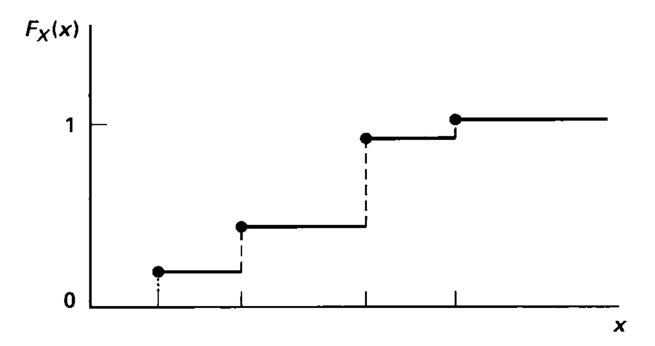


Figure 2.5-1 The cumulative distribution function for a discrete random variable.

$$\begin{aligned} P_X(x) &= P[X = x] \\ &= P[X \le x] - P[X < x]. \end{aligned} \tag{2.5-5}$$

Thus, $P_X(x) = 0$ everywhere where $F_X(x)$ is continuous and has nonzero values only where there is a discontinuity, that is, jump, in the CDF. If we denote $P[X < x]$ by $F_X(x^-)$, then at the jumps x_i, $i = 1, 2, \ldots$, the finite values of $P_X(x_i)$ can be computed from $P_X(x_i) = F_X(x_i) - F_X(x_i^-)$.

The probability mass function is used when there are at most a countable set of outcomes of the random experiment. Indeed $P_X(x_i)$ lends itself to the following frequency interpretation: Perform an experiment n times and let n_i be the number of tries that x_i appears as an outcome. Then, for n large,

$$P_X(x_i) \simeq \frac{n_i}{n}. \tag{2.5-6}$$

Because the PMF is so closely related to the frequency notion of probability, it is sometimes called the *frequency function.*

Since for a discrete RV $F_X(x)$ is not continuous $f_X(x)$, strictly speaking, does not exist. Nevertheless, with the introduction of Dirac delta functions,† we shall be able to assign pdf's to discrete RVs as well. The CDF for a discrete RV is given by

$$F_X(x) \triangleq P[X \le x] = \sum_{\text{all } x_i \le x} P_X(x_i) \tag{2.5-7}$$

and, more generally, for any event B when X is discrete:

$$P[B] = \sum_{\text{all } x_i \in B} P_X(x_i). \tag{2.5-8}$$

†Also called impulses or impulse functions. Named after the English physicist Paul A. M. Dirac (1902–1984). Delta functions are discussed in Section B.2 of Appendix B.

Some Common Discrete Random Variables

1. *Bernoulli* random variable B with parameter p $(0 < p < 1, q \triangleq 1 - p)$:

$$P_B(k) = \begin{cases} q, \; k = 0, \\ p, \; k = 1, \\ 0, \;\; \text{else}, \end{cases} \tag{2.5-9}$$

$$= q\delta(k) + p\delta(k-1), \quad \text{by use of discrete delta function}^\dagger \; \delta(k). \tag{2.5-10}$$

The Bernoulli random variable appears in those situations where the outcome is one of two possible states, for example, whether a particular bit in a digital sequence is "one" or "zero." The Bernoulli PMF can be conveniently written as $P_B(k) = p^k q^{1-k}$ for $k = 0$ or 1 and then zero elsewhere. The corresponding CDF is given as

$$F_B(k) = \begin{cases} 0, \; k < 0, \\ q, \; k = 0, \\ 1, \; k \geq 1. \end{cases}$$

$$= qu(k) + pu(k-1) \quad \text{by use of unit-step function} \;\; u(k).$$

2. *Binomial* random variable K with parameters n and p $(n = 1, 2, \ldots; 0 < p < 1)$ and k an integer:

$$P_K(k) = \begin{cases} \binom{n}{k} p^k q^{n-k}, \; 0 \leq k \leq n, \\ \quad 0, \qquad\qquad \text{else}, \end{cases} \tag{2.5-11}$$

$$= \binom{n}{k} p^k q^{n-k} \left[u(k) - u(n-k) \right]. \tag{2.5-12}$$

The binomial random variable appears in games of chance, military defense strategies, failure analysis, and many other situations. Its corresponding CDF is given as $(l, k, n$ are integers)

$$F_K(k) = \begin{cases} \qquad 0, & k < 0, \\ \sum_{l=0}^{k} \binom{n}{l} p^l q^{n-l}, & 0 \leq k < n, \\ \qquad 1, & k \geq n. \end{cases}$$

3. *Poisson* random variable X with parameter $\mu (> 0)$ and k an integer:

$$P_X(k) = \begin{cases} \frac{\mu^k}{k!} e^{-\mu}, \; 0 \leq k < \infty, \\ \quad 0, \qquad\quad \text{else}. \end{cases} \tag{2.5-13}$$

The Poisson law is widely used in every branch of science and engineering (see Section 1.10). We can write the Poisson PMF in a single line by use of the unit-step function $u(k)$ as

$$P_X(k) = \frac{\mu^k}{k!} e^{-\mu} u(k),$$

† Recall that the discrete delta function has value 1 when the argument is 0 and has value 0 for every other value.

where the discrete unit-step function is defined by

$$u(k) \triangleq \begin{matrix} 1, & 0 \leq k < \infty, \\ 0, & -\infty < k < 0. \end{matrix}$$

4. *Geometric* random variable K with parameters $p > 0, q > 0, (p + q = 1)$ and k an integer:

$$P_K(k) = \begin{cases} pq^k, & 0 \leq k < \infty, \\ 0, & \text{else}, \end{cases}$$
$$= pq^k u(k).$$

The corresponding CDF is given by a finite sum of the geometric series (ref. Appendix A) as

$$F_K(k) = \begin{cases} 0, & k < 0, \\ p\left(\frac{1-q^{k+1}}{1-q}\right), & 0 \leq k < \infty, \end{cases}$$
$$= p\left(\frac{1-q^{k+1}}{1-q}\right)u(k).$$

This distribution[†] was first seen in Example 1.9-4. As there, also note the variant $pq^{n-1}, n \geq 1$, also called geometric RV.

Example 2.5-1

(CDF *of Poisson RV*) Calculating the CDF of a Poisson random variable proceeds as follows. Let X be a Poisson random variable with parameter μ (>0). Then by definition the PMF is $P_X(k) = \frac{\mu^k}{k!}e^{-\mu}u(k)$. Then the CDF $F_X(k) = 0$ for $k < 0$. For $k \geq 0$, we have

$$F_X(k) = \sum_{l=0}^{k} \frac{\mu^l}{l!}e^{-\mu}$$
$$= \left(\sum_{l=0}^{k} \frac{\mu^l}{l!}\right)e^{-\mu}.$$

Table 2.5-1 lists the common discrete RVs, their PMFs, and their CDFs.

Sometimes an RV is neither purely discrete nor purely continuous. We call such an RV a *mixed RV*. The CDF of a mixed RV is shown in Figure 2.5-2. Thus, $F_X(x)$ is discontinuous but not a staircase-type function.

[†]Note that we sometimes speak of the probability *distribution* in a general sense without meaning the distribution function per se. Here we give a PMF to illustrate the geometric distribution.

Table 2.5-1 Table Common Discrete RVs, PMFs, and CDFs

Family	PMF $P_K(k)$	CDF $F_K(k)$
Bernoulli p, q	$q\delta(k) + p\delta(k-1)$	$qu(k) + pu(k-1)$
Binomial n, k	$\binom{n}{k} p^k q^{n-k} [u(k) - u(n-k)]$	$\begin{cases} 0, & k < 0, \\ \sum_{l=0}^{k} \binom{n}{l} p^l q^{n-l}, & 0 \leq k < n \\ 1, & k \geq n. \end{cases}$
Poisson $\mu > 0$	$\frac{\mu k}{k!} e^{-\mu} u(k)$	$\frac{\gamma(k+1, \mu)^\dagger}{k!} \times u(k)$
Geometric p, q	$pq^k u(k)$	$p\left(\frac{1-q^{k+1}}{1-q}\right) u(k)$

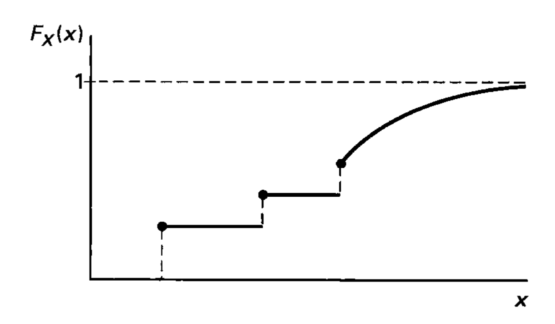


Figure 2.5-2 The CDF of a mixed RV.

The distinction between continuous and discrete RVs is somewhat artificial. Continuous and discrete RVs are often regarded as different objects even though the only real difference between them is that for the former the CDF is continuous while for the latter it is not. By introducing delta functions we can, to a large extent, treat them in the same fashion and compute probabilities for both continuous and discrete RVs by integrating pdf's.

Returning now to Equation 2.5-7, which can be written as

$$F_X(x) = \sum_{i=-\infty}^{\infty} P_X(x_i) u(x - x_i), \tag{2.5-14}$$

and using the results from the section on delta functions in Appendix B enables us to write for a discrete RV

$$f_X(x) = \frac{dF_X(x)}{d_X} = \sum_{i=-\infty}^{\infty} P_X(x_i) \delta(x - x_i). \tag{2.5-15}$$

†See Appendix B for a definition of the incomplete gamma.

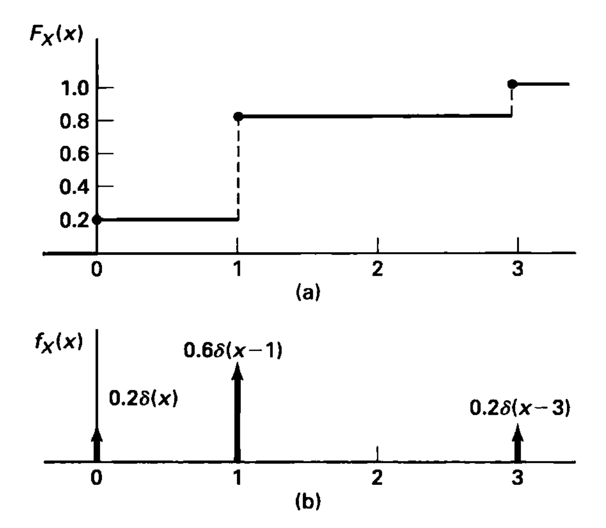


Figure 2.5-3 (a) CDF of a discrete RV X; (b) pdf of X using delta functions.

Example 2.5-2

(*practice example*) Let X be a discrete RV with distribution function as shown in Figure 2.5-3(a). The pdf of X is

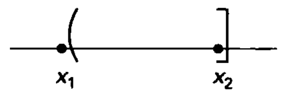
$$f_X(x) = \frac{dF_X}{dx} = 0.2\delta(x) + 0.6\delta(x-1) + 0.2\delta(x-3)$$

and is shown in Figure 2.5-3(b). To compute probabilities from the pdf for a discrete RV, great care must be used in choosing the interval of integration. Thus,

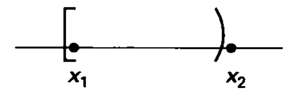
$$F_X(x) = \int_{-\infty}^{x^+} f_X(\xi)d\xi,$$

which includes the delta function at x if there is one there.

Similarly $P[x_1 < X \leq x_2]$ involves the interval



and includes the impulse at x_2 (if there is one there) but excludes what happens at x_1. On the other hand $P[x_1 \leq X < x_2]$ involves the interval



and therefore

$$P(x_1 \le X < x_2) = \int_{x_1^-}^{x_2^-} f_X(\xi)d\xi.$$

Applied to the foregoing example, these formulas give

$$P[X \le 1.5] = F_X(1.5) = 0.8 \tag{2.5-16}$$

$$P[1 < X \le 3] = 0.2 \tag{2.5-17}$$

$$P[1 \le X < 3] = 0.6. \tag{2.5-18}$$

Example 2.5-3

(*Practice example*) The pdf associated with the Poisson law with parameter a is

$$f_X(x) = e^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!} \delta(x-k).$$

Example 2.5-4

(*Practice example*) The pdf associated with the binomial law $b(k; n, p)$ is

$$f_X(x) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \delta(x-k).$$

Example 2.5-5

(*Practice example*) The pdf of a mixed RV is shown in Figure 2.5-4. (1) What is the constant K? (2) Compute $P[X \le 5]$, $P[5 \le X < 10]$. (3) Draw the distribution function.

Solution (1) Since

$$\int_{-\infty}^{\infty} f_X(\xi)d\xi = 1,$$

we obtain $10K + 0.25 + 0.25 = 1 \Rightarrow K = 0.05$.

(2) Since $P[X \le 5] = P[X < 5] + P[X = 5]$, the impulse at $x = 5$ must be included. Hence

$$P[X \le 5] = \int_0^{5^+} [0.05 + 0.25\delta(\xi - 5)]d\xi = 0.5.$$

To compute $P(5 \le X < 10)$, we leave out the impulse at $x = 10$ but include the impulse at $x = 5$. Thus,

$$P[5 \le X < 10] = \int_{5^-}^{10^-} [0.05 + 0.25\delta(\xi - 5)]d\xi = 0.5.$$

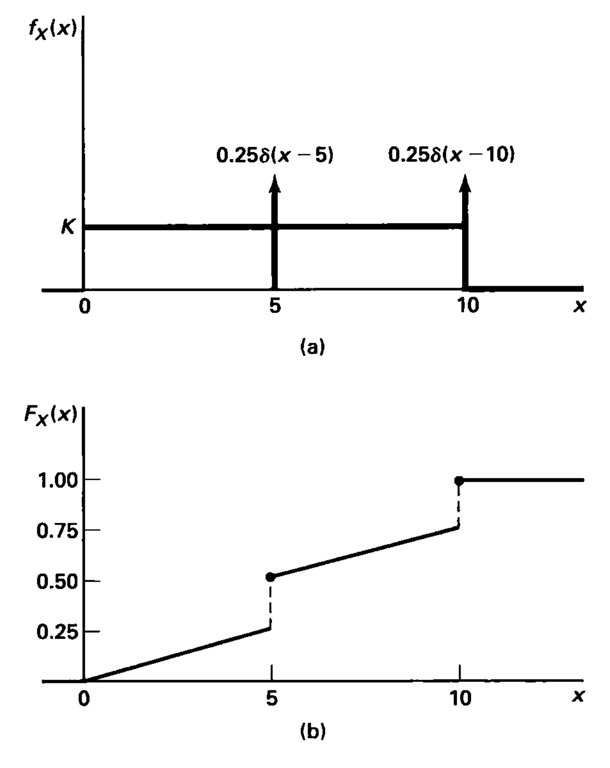


Figure 2.5-4 (a) pdf of a mixed RV for Example 2.5-5; (b) computed pdf.

2.6 CONDITIONAL AND JOINT DISTRIBUTIONS AND DENSITIES

Consider the event C consisting of all outcomes $\zeta \in \Omega$ such that $X(\zeta) \leq x$ and $\zeta \in B \subset \Omega$, where B is another event. Then, by definition, the event C is the set intersection of the two events $\{\zeta\colon X(\zeta) \leq x\}$ and $\{\zeta\colon \zeta \in B\}$. We define the *conditional distribution function of X given the event B as*

$$F_X(x|B) \triangleq \frac{P[C]}{P[B]} = \frac{P[X \leq x, B]}{P[B]}, \tag{2.6-1}$$

where $P[X \leq x, B]$ is the probability of the joint event $\{X \leq x\} \cap B$ and $P[B] \neq 0$. If $x = \infty$, the event $\{X \leq \infty\}$ is the certain event Ω and since $\Omega \cap B = B$, $F_X(\infty|B) = 1$. Similarly, if $x = -\infty$, $\{X \leq -\infty\} = \phi$ and since $\Omega \cap \phi = \phi$, $F_X(-\infty|B) = 0$. Continuing in this fashion, it is not difficult to show that $F_X(x|B)$ has all the properties of an ordinary distribution, that is, $x_1 \leq x_2 \rightarrow F_X(x_1|B) \leq F_X(x_2|B)$.

For example, consider the event $\{X \leq x_2, B\}$ and write (assuming $x_2 \geq x_1$)

$$\{X \leq x_2, B\} = \{X \leq x_1, B\} \cup \{x_1 < X \leq x_2, B\}.$$

Since the two events on the right are disjoint, their probabilities add and we obtain

$$P[X \leq x_2, B] = P[X \leq x_1, B] + P[x_1 < X \leq x_2, B]$$

or

$$P[X \leq x_2|B]P[B] = P[X \leq x_1|B]P[B] + P[x_1 < X \leq x_2|B]P[B].$$

Thus when $P[B] \neq 0$, we obtain after rearranging terms and dividing by B

$$\begin{aligned} P[x_1 < X \leq x_2|B] &= P[X \leq x_2|B] - P[X \leq x_1|B] \\ &= F_X(x_2|B) - F_X(x_1|B). \end{aligned} \tag{2.6-2}$$

Generally the event B will be expressed on the probability space $(R, \mathscr{B}, P_X)$ rather than the original space $(\Omega, \mathscr{F}, P)$. The conditional pdf is simply

$$f_X(x|B) \triangleq \frac{dF_X(x|B)}{dx}. \tag{2.6-3}$$

Following are some examples.

Example 2.6-1

(*evaluating conditional* CDFs) Let $B \triangleq \{X \leq 10\}$. We wish to compute $F_X(x|B)$.

(i) For $x \geq 10$, the event $\{X \leq 10\}$ is a subset of the event $\{X \leq x\}$. Hence $P[X \leq 10, X \leq x] = P[X \leq 10]$ and use of Equation 2.6-1 gives

$$F_X(x|B) = \frac{P[X \leq x, X \leq 10]}{P[X \leq 10]} = 1.$$

(ii) For $x \leq 10$, the event $\{X \leq x\}$ is a subset of the event $\{X \leq 10\}$. Hence $P[X \leq 10, X \leq x] = P[X \leq x]$ and

$$F_X(x|B) = \frac{P[X \leq x]}{P[X \leq 10]}.$$

The result is shown in Figure 2.6-1. We leave as an exercise to the reader to compute $F_X(x|B)$ when $B = \{b < X \leq a\}$.

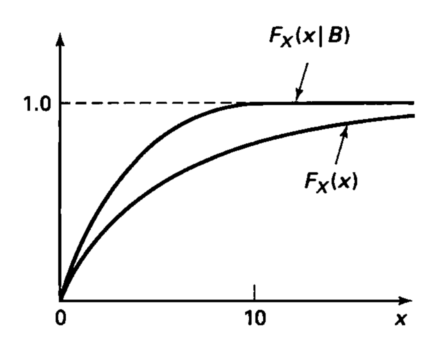


Figure 2.6-1 Conditional and unconditional CDFs of X.

Example 2.6-2

(*Poisson conditioned on even*) Let X be a Poisson RV with parameter $\mu\,(>0)$. We wish to compute the conditional PMF and CDF of X given the event $\{X = 0, 2, 4, \ldots\} \triangleq \{X(\text{is})\text{ even}\}$. First observe that $P[X\text{ even}]$ is given by

$$P[X = 0, 2, \ldots] = \sum_{k=0,2,\ldots}^{\infty} \frac{\mu^k}{k!} e^{-\mu}.$$

Then for X odd, we have

$$P[X = 1, 3, \ldots] = \sum_{k=1,3,\ldots}^{\infty} \frac{\mu^k}{k!} e^{-\mu}.$$

From these relations, we obtain

$$\begin{aligned}\sum_{k\geq 0\text{ and even}} \frac{\mu^k}{k!} e^{-\mu} - \sum_{k\geq 0\text{ and odd}} \frac{\mu^k}{k!} e^{-\mu} &= \sum_{k=0}^{\infty} \frac{\mu^k}{k!} (-1)^k e^{-\mu} \\ &= \sum_{k=0}^{\infty} \frac{(-\mu)^k}{k!} e^{-\mu} \\ &= e^{-\mu} e^{-\mu} \\ &= e^{-2\mu}\end{aligned}$$

and

$$\sum_{k\geq 0\text{ and even}} \frac{\mu^k}{k!} e^{-\mu} + \sum_{k\geq 0\text{ and odd}} \frac{\mu^k}{k!} e^{-\mu} = 1.$$

Hence $P[X\text{ even}] = P[X = 0, 2, \ldots] = \frac{1}{2}(1+e^{-2\mu})$. Using the definition of conditional PMF, we obtain

$$P_X(k|X\text{ even}) = \frac{P[X = k, X\text{ even}]}{P[X\text{ even}]}.$$

If k is even, then $\{X = k\}$ is a subset of $\{X \text{ even}\}$. If k is odd, $\{X = k\} \cap \{X \text{ even}\} = \phi$. Hence $P[X = k, X \text{ even}] = P[X = k]$ for k even and it equals 0 for k odd. So we have

$$P_X(k|X \text{ even}) = \begin{cases} \frac{2}{(1+2e^{-\mu})}\frac{\mu^k}{k!}e^{-\mu}, & k \geq 0 \text{ and even}, \\ 0, & k \text{ odd}. \end{cases}$$

The conditional CDF is then

$$\begin{aligned} F_X(x|X \text{ even}) &= \sum_{\text{all } k \leq x} P_X(k|X \text{ even}) \\ &= \sum_{\substack{0 \leq k \leq x \\ \text{and even}}} \frac{2}{(1+2e^{-\mu})}\frac{\mu^k}{k!}e^{-\mu}. \end{aligned}$$

Let us next derive some important formulas involving conditional CDFs and pdf's.

The distribution function written as a weighted sum of conditional distribution functions. Equation 1.6-7 in Chapter 1 gave the probability of the event B in terms of n mutually exclusive and exhaustive events $\{A_i\}, i = 1, \ldots, n$, defined on the same probability space as B. With $B \triangleq \{X \leq x\}$, we immediately obtain from Equation 1.6-7:

$$F_X(x) = \sum_{i=1}^{n} F_X(x|A_i)P[A_i]. \tag{2.6-4}$$

Equation 2.6-4 describes $F_X(x)$ as a weighted sum of conditional distribution functions. One way to view Equation 2.6-4 is an "average" over all the conditional CDFs.† Since we haven't yet made concrete the notion of average (this will be done in Chapter 4), we ask only that the reader accept the nomenclature since it is in use in the technical literature.

Example 2.6-3

(*defective memory chips*) In the automated manufacturing of computer memory chips, company Z produces one defective chip for every five good chips. The defective chips (DC) have a time of failure X that obeys the CDF

$$F_X(x|DC) = (1 - e^{-x/2})u(x) \quad (x \text{ in months})$$

while the time of failure for the good chips (GC) obeys the CDF

$$F_X(x|GC) = (1 - e^{-x/10})u(x) \quad (x \text{ in months}).$$

The chips are visually indistinguishable. A chip is purchased. What is the probability that the chip will fail before six months of use?

---

†For this reason, when $F_X(x)$ is written as in Equation 2.6-4, it is sometimes called the *average* distribution function.

Solution The unconditional CDF for the chip is, from Equation 2.6-4,

$$F_X(x) = F_X(x|DC)P[DC] + F_X(x|GC)P[GC],$$

where $P[DC]$ and $P[GC]$ are the probabilities of selecting a defective and good chip, respectively. From the given data $P[DC] = 1/6$ and $P[GC] = 5/6$. Thus,

$$\begin{aligned} F_X(6) &= [1 - e^{-3}]\tfrac{1}{6} + [1 - e^{-0.6}]\tfrac{5}{6} \\ &= 0.158 + 0.376 = 0.534. \end{aligned}$$

Bayes' formula for probability density functions. Consider the events B and $\{X = x\}$ defined on the same probability space. Then from the definition of conditional probability, it seems reasonable to write

$$P[B|X = x] = \frac{P[B, X = x]}{P[X = x]}. \tag{2.6-5}$$

The problem with Equation 2.6-5 is that if X is a continuous RV, then $P[X = x] = 0$. Hence Equation 2.6-5 is undefined. Nevertheless, we can compute $P[B|X = x]$ by taking appropriate limits of probabilities involving the event $\{x < X \leq x + \Delta x\}$. Thus, consider the expression

$$P[B|x < X \leq x + \Delta x] = \frac{P[x < X \leq x + \Delta x|B]P[B]}{P[x < X \leq x + \Delta x]}.$$

If we (i) divide numerator and denominator of the expression on the right by Δx, (ii) use the fact that $P[x < X \leq x + \Delta x|B] = F(x + \Delta x|B) - F(x|B)$, and (iii) take the limit as $\Delta x \to 0$, we obtain

$$\begin{aligned} P[B|X = x] &= \lim_{\Delta x \to 0} P[B|x < X \leq x + \Delta x] \\ &= \frac{f_X(x|B)P[B]}{f_X(x)}, \qquad f(x) \neq 0. \end{aligned} \tag{2.6-6}$$

The quantity on the left is sometimes called the *a posteriori* probability (or *a posteriori* density) of B given $X = x$. Multiplying both sides of Equation 2.6-6 by $f_X(x)$ and integrating enables us to obtain the important result

$$P[B] = \int_{-\infty}^{\infty} P[B|X = x]f_X(x)dx. \tag{2.6-7}$$

In line with the terminology used in this section, $P[B]$ is sometimes called the *average probability* of B, the usage being suggested by the form of Equation 2.6-7.

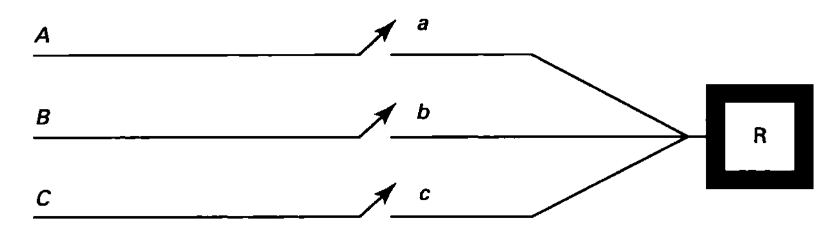


Figure 2.6-2 Based upon observing the signal, the receiver R must decide which switch was closed or, equivalently, which of the sources A, B, C was responsible for the signal. Only one switch can be closed at the time the receiver is on.

Example 2.6-4

(*detecting closed switch*) A signal, X, can come from one of three different sources designated as A, B, or C. The signal from A is distributed as $N(-1,4)$; the signal from B is distributed as $N(0,1)$; and the signal from C has an $N(1,4)$ distribution. In order for the signal to reach its destination at R, the switch in the line must be closed. Only one switch can be closed when the signal X is observed at R, but it is not known which switch it is. However, it is known that switch a is closed twice as often as switch b, which is closed twice as often as switch c (Figure 2.6-2).

(a) Compute $P[X \leq -1]$;
(b) Given that we observe the event $\{X > -1\}$, from which source was this signal most likely?

Solution (a) Let $P[A]$ denote the probability that A is responsible for the observation at R, that is, switch a is closed. Likewise for $P[B]$, $P[C]$. Then from the information about the switches we get $P[A] = 2P[B] = 4P[C]$ and $P[A]+P[B]+P[C] = 1$. Hence $P[A] = 4/7$, $P[B] = 2/7$, $P[C] = 1/7$. Next we compute $P[X \leq -1]$ from

$$P[X \leq -1] = P[X \leq -1|A]P[A] + P[X \leq -1|B]P[B] + P[X \leq -1|C]P[C],$$

where

$$P[X \leq -1|A] = 1/2 \tag{2.6-8}$$

$$P[X \leq -1|B] = 1/2 - \text{erf}(1) = 0.159 \tag{2.6-9}$$

$$P[X \leq -1|C] = 1/2 - \text{erf}(1) = 0.159. \tag{2.6-10}$$

Hence $P[X \leq -1] = 1/2 \times 4/7 + 0.159 \times 2/7 + 0.159 \times 1/7 \approx 0.354$.

(b) We wish to compute $\max\{P[A|X > -1], P[B|X > -1], P[C|X > -1]\}$. To enable this computation, we note that $P[X > -1|A] = 1 - P[X \leq -1|A]$, and so on, for B and C. Concentrating on source A, and using Bayes' rule, we get

$$P[A|X > -1] = \frac{\{1 - P[X \leq -1|A]\} \times P[A]}{1 - P[X \leq -1]},$$

which, using the values already computed, yields $P[A|X > -1] = 0.44$.

Repeating the calculation for the other sources, we obtain

$$P[B|X > -1] = 0.372, \tag{2.6-11}$$

$$P[C|X > -1] = 0.186. \tag{2.6-12}$$

Hence, since the maximum *a posteriori* probability favors A, source A was the most likely cause of the event $\{X > -1\}$.

Poisson transform. An important specific example of Equation 2.6-7 is the so-called *Poisson transform* in which B is the event that a random variable Y takes on an integer value k from the set $\{0,1,\ldots,\}$ that is, $B \triangleq \{Y = k\}$ and X is the Poisson parameter, treated here as a random variable with pdf $f_X(x)$. The ordinary Poisson law

$$P[Y = k] = e^{-\mu}\frac{\mu^k}{k!}, \qquad k \geq 0, \tag{2.6-13}$$

where μ is the average number of events in a given interval (time, distance, volume, and so forth), treats the parameter μ as a constant. But in many situations the underlying phenomenon that determines μ is itself random and μ must be viewed as a random outcome, that is, the outcome of a random experiment. Thus, there are two elements of randomness: the random value of μ and the random outcome $\{Y = k\}$. When μ is random it seems appropriate to replace it by the notation of a random variable, say X. Thus, for any given outcome $\{X = x\}$ the probability $P[Y = k|X = x]$ is Poisson; but the *unconditional* probability of the event $\{Y = k\}$ is not necessarily Poisson. Because both the number of events and the Poisson parameter are random, this situation is sometimes called *doubly stochastic*. From Equation 2.6-7 we obtain for the unconditional PMF of Y

$$P_Y(k) = \int_0^\infty \frac{x^k}{k!}e^{-x}f_X(x)dx, \qquad k \geq 0. \tag{2.6-14}$$

The above Equation is known as the Poisson transform and can be used to obtain $f_X(x)$ if $P_Y(k)$ is obtained by experimentation. The mechanism by which $f_X(x)$ is obtained from $P_Y(k)$ is the *inverse Poisson transform*. The derivation of the latter is as follows. Let

$$F(\omega) \triangleq \frac{1}{2\pi}\int_0^\infty e^{j\omega x}e^{-x}f_X(x)dx, \tag{2.6-15}$$

that is, the inverse Fourier transform of $e^{-x}f_X(x)$. Since

$$e^{j\omega x} = \sum_{k=0}^{\infty}[j\omega x]^k/k!, \tag{2.6-16}$$

we obtain

$$F(\omega) = \frac{1}{2\pi}\sum_{k=0}^{\infty}(j\omega)^k\int_0^\infty \frac{x^k}{k!}e^{-x}f_X(x)dx$$

$$= \frac{1}{2\pi} \sum_{k=0}^{\infty} j\omega^k P_Y(k) \tag{2.6-17}$$

Thus, $F(\omega)$ is known if $P_Y(k)$ is known, Taking the forward Fourier transforms of $F(\omega)$ yields

$$e^{-x} f_X(\omega) = e^x \int_{-\infty}^{\infty} F(\omega) e^{-j\omega x} d\omega.$$

or

$$f_X(x) = e^x \int_{-\infty}^{\infty} F(\omega) e^{-j\omega x} d\omega. \tag{2.6-18}$$

Equation 2.6-18 is the inverse relation we have been seeking. Thus to summarize: If we know $P_Y(k)$, we can compute $F(\omega)$. Knowing $F(\omega)$ enables us to obtain $f_X(x)$ by a Fourier transform. We illustrate the Poisson transform with an application from optical communication theory.

Example 2.6-5

(*optical communications*) In an optical communication system, light from the transmitter strikes a photodetector, which generates a photocurrent consisting of valence electrons having become conduction electrons (Figure 2.6-3).

It is known from physics that if the transmitter uses coherent laser light of constant intensity the Poisson parameter X has pdf

$$f_X(x) = \delta(x - x_0) \qquad x_o > 0, \tag{2.6-19}$$

where x_o, except for a constant, is the laser intensity. On the other hand, if the transmitter uses thermal illumination, then the Poisson parameter X obeys the exponential law:

$$f_X(x) = \frac{1}{\mu} e^{-(1/\mu)x} u(x), \tag{2.6-20}$$

where $\mu > 0$ is now just a parameter, but one that will later be shown to be the true mean value of X. Compute the PMF for the electron-count variable Y.

Solution For coherent laser illumination we obtain from Equation 2.6-14

$$P_Y(k) = \int_0^{\infty} \frac{x^k}{k!} e^{-x} \delta(x - x_0) dx \tag{2.6-21}$$

$$= \frac{x_o^k}{k!} e^{-x_0}, k \geq 0. \tag{2.6-22}$$

Thus, for coherent laser illumination, the photoelectrons *obey* the Poisson law. For thermal illumination, we obtain

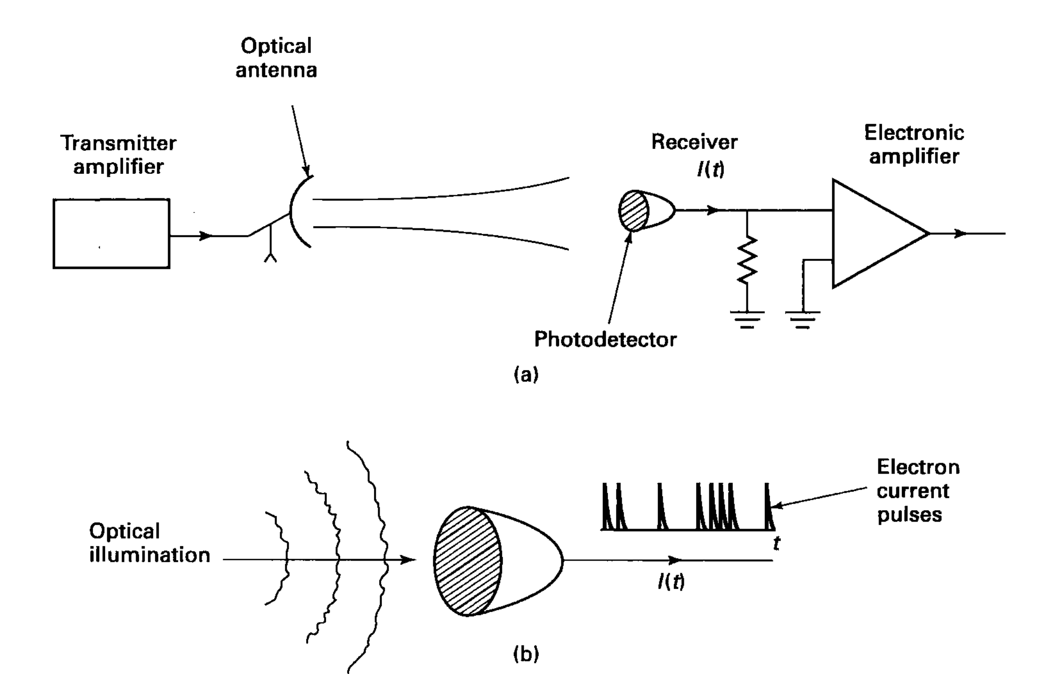


Figure 2.6-3 (a) Optical communication system; (b) output current from photodetector.

$$
\begin{aligned}
P_Y(k) &= \int_0^\infty \frac{x^k}{k!}e^{-x}\frac{1}{\mu}e^{-x/\mu}dx \\
&= \frac{1}{\mu k!}\int_0^\infty x^k e^{-x/\alpha}dx, \quad \text{with} \quad \alpha \triangleq \frac{\mu}{\mu+1}, \\
&= \frac{\alpha^{k-1}}{\mu k!}\int_0^\infty z^k e^{-z}dz, \quad \text{with} \quad z \triangleq x/\alpha, \\
&= \frac{\alpha^{k-1}}{\mu k!}\Gamma(k+1), \quad \text{where } \Gamma \text{ denotes the Gamma function (see Appendix B)}, \\
&= \frac{\alpha^{k-1}}{\mu k!}k! \\
&= \frac{\alpha^{k-1}}{\mu} \\
&= \frac{\mu^k}{[1+\mu]^{k+1}}, k \geq 0.
\end{aligned}
\tag{2.6-23}
$$

This PMF law is known as the *geometric* distribution and is sometimes called *Bose–Einstein statistics* [2-4]. It obeys the interesting recurrence relation

$$P_Y(k+1) = \frac{\mu}{1+\mu} P_Y(k). \tag{2.6-24}$$

Depending on which illumination applies, the statistics of the photocurrents are widely dissimilar.

Joint distributions and densities. As stated in Section 2.1, it is possible to define more than one random variable on a probability space. For example, consider a probability space $(\Omega, \mathscr{F}, P)$ involving an underlying experiment consisting of the *simultaneous* throwing of two fair coins. Here the ordering is not important and the only elementary outcomes are ζ_1 =HH, ζ_2 =HT, ζ_3 =TT, the sample space is $\Omega = \{\text{HH}, \text{HT}, \text{TT}\}$, the σ-field of events is ϕ, Ω, {HT}, {TT}, {HH}, {TT *or* HT}, {HH *or* HT}, and {HH *or* TT}. The probabilities are easily computed and are, respectively, 0, 1, 1/2, 1/4, 1/4, 3/4, 3/4, and 1/2. Now define two random variables

$$X_1(\zeta) = \begin{cases} 0, & \textit{if at least one H} \\ 1, & \textit{otherwise} \end{cases} \tag{2.6-25}$$

$$X_2(\zeta) = \begin{cases} -1, & \textit{if one H and one T} \\ +1, & \textit{otherwise.} \end{cases} \tag{2.6-26}$$

Then $P[X_1 = 0] = 3/4$, $P[X_1 = 1] = 1/4$, $P[X_2 = -1] = 1/2$, $P[X_2 = 1] = 1/2$. Also we can easily compute the probability of joint events, for example, $P[X_1 = 0, X_2 = 1] = P[\{\text{HH}\}] = 1/4$.

In defining more than one random variable on a probability space, it is possible to define degenerate random variables. For example suppose the underlying experiment consists of observing the number ζ that is pointed to when a spinning wheel, numbered 0 to 100, comes to rest. Suppose we let $X_1(\zeta) = \zeta$ and $X_2(\zeta) = e^{\zeta}$. This situation is degenerate because observing one random variable completely specifies the other. In effect the uncertainty is associated with only one random variable, not both; we might as well forget about observing the other one. If we define more than one random variable on a probability space, degeneracy can be avoided if the underlying experiment is complex enough, or rich enough in outcomes. In the example we considered at the beginning, observing that $X_1 = 0$ doesn't specify the value of X_2 while observing $X_2 = 1$ doesn't specify the value of X_1.

The event $\{X \le x,\ Y \le y\} \triangleq \{X \le x\} \cap \{Y \le y\}$ consists of all outcomes $\zeta \in \Omega$ such that $X(\zeta) \le x$ and $Y(\zeta) \le y$. The point set induced by the event $\{X \le x, Y \le y\}$ is the shaded region in the $x'y'$ plane shown in Figure 2.6-4. In the diagram the numbers x, y are shown positive. In general they can have any value. The *joint cumulative distribution function* of X and Y is defined by

$$F_{XY}(x,y) = P[X \le x, Y \le y]. \tag{2.6-27}$$

By definition $F_{XY}(x,y)$ is a probability; thus it follows that $F_{XY}(x,y) \ge 0$ for all x, y. Since $\{X \le \infty, Y \le \infty\}$ is the certain event, $F_{XY}(\infty,\infty) = 1$. The point set associated with the certain event is the whole $x'y'$ plane. The event $\{X \le -\infty, y \le -\infty\}$ is the

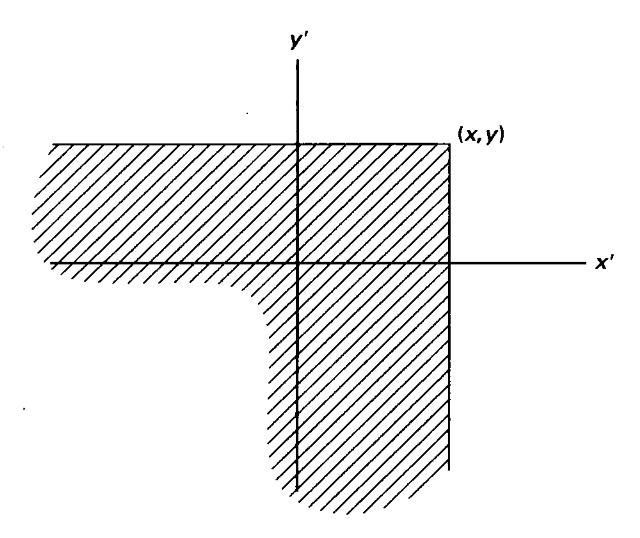


Figure 2.6-4 Point set associated with the event $\{X \leq x, Y \leq y\}$.

impossible event and therefore $F_{XY}(-\infty, -\infty) = 0$. The reader should consider the events $\{X \leq x, Y \leq -\infty\}$ and $\{X \leq -\infty, Y \leq y\}$; are they impossible events also?

Since $\{X \leq \infty\}$ and $\{Y \leq \infty\}$ are certain events, and for any event B, $B \cap \Omega = B$, we obtain

$$\begin{aligned}\{X \leq x, Y \leq \infty\} &= \{X \leq x\} \cap \{Y \leq \infty\} \\ &= \{X \leq x\} \cap \Omega \\ &= \{X \leq x\}\end{aligned} \tag{2.6-28}$$

so that

$$F_{XY}(x, \infty) = F_X(x) \tag{2.6-29a}$$

$$F_{XY}(\infty, y) = F_Y(y) \tag{2.6-29b}$$

If $F_{XY}(x, y)$ is continuous and differentiable, the joint pdf can be obtained from

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x\, \partial y}[F_{XY}(x, y)]. \tag{2.6-30}$$

It follows then, that

$$f_{XY}(x, y) dx\, dy = P[x < X \leq x + dx, y < Y \leq y + dy]$$

and hence that $f_{XY}(x, y) \geq 0$ for all (x, y).

By twice integrating Equation 2.6-30, we obtain

$$F_{XY}(x, y) = \int_{-\infty}^{x} d\xi \int_{-\infty}^{y} d\eta f_{XY}(\xi, \eta). \tag{2.6-31}$$

Equation 2.6-31 says that $F_{XY}(x,y)$ is the integral of the nonnegative function $f_{XY}(x,y)$ over the surface shown in Figure 2.6-4. It follows that integrating $f_{XY}(x,y)$ over a larger surface will generally yield a larger probability (never a smaller one!) than integrating over a smaller surface. From this we can deduce some obvious but important results. Thus, if (x_1,y_1) and (x_2,y_2) denote two pairs of numbers and if $x_1 \leq x_2$, $y_1 \leq y_2$, then $F_{XY}(x_1,y_1) \leq F_{XY}(x_2,y_2)$. In general, $F_{XY}(x,y)$ increases as (x,y) moves up and to the right and decreases as (x,y) moves down and to the left. Also F_{XY} is continuous from above and from the right, that is, at a point of discontinuity, say x_0, y_0, with ε, $\delta > 0$:

$$F_{XY}(x_0,y_0) = \lim_{\substack{\varepsilon \to 0 \\ \delta \to 0}} F_{XY}(x_0+\varepsilon, y_0+\delta).$$

Thus, at a point of discontinuity, F_{XY} assumes the value immediately to the right and above the point.

Properties of joint CDF $F_{XY}(x,y)$

(i) $F_{XY}(\infty,\infty) = 1$; $F_{XY}(-\infty,y) = F_{XY}(x,-\infty) = 0$; also $F_{XY}(x,\infty) = F_X(x)$; $F_{XY}(\infty,y) = F_Y(y)$.

(ii) If $x_1 \leq x_2$, $y_1 \leq y_2$, then $F_{XY}(x_1,y_1) \leq F_{XY}(x_2,y_2)$.

(iii) $F_{XY}(x,y) = \lim_{\substack{\varepsilon \to 0 \\ \delta \to 0}} F_{XY}(x+\varepsilon, y+\delta)$ $\quad \varepsilon, \delta > 0$ (continuity from the right and from above).

(iv) For all $x_2 \geq x_1$ and $y_2 \geq y_1$, we must have

$$F_{XY}(x_2,y_2) - F_{XY}(x_2,y_1) - F_{XY}(x_1,y_2) + F_{XY}(x_1,y_1) \geq 0.$$

This last and key property (iv) is a two-dimensional generalization of the nondecreasing property for one-dimensional CDFs, that is, $F_X(x_2) - F_X(x_1) \geq 0$ for all $x_2 \geq x_1$. It arises out of the need for the event $\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}$ to have nonnegative probability. The point set induced by this event is shown in Figure 2.6-5.

The key to this computation is to observe that the set $\{X \leq x_2, Y \leq y_2\}$ lends itself to the following decomposition into disjoint sets:

$$\begin{aligned}\{X \leq x_2, Y \leq y_2\} = &\{x_1 < X \leq x_2, y_1 < Y \leq y_2\} \\ &\cup \{x_1 < X \leq x_2, Y \leq y_1\} \cup \{X \leq x_1, y_1 < Y \leq y_2\} \\ &\cup \{X \leq x_1, Y \leq y_1\}. \end{aligned} \tag{2.6-32}$$

Now using the induced result from Axiom 3 (Equation 1.5-3), we obtain

$$\begin{aligned}F_{XY}(x_2,y_2) = &P[x_1 < X \leq x_2, y_1 < Y \leq y_2] \\ &+ P[x_1 < X \leq x_2, Y \leq y_1] + P[X \leq x_1, y_1 < Y \leq y_2] \\ &+ F_{XY}(x_1,y_1). \end{aligned} \tag{2.6-33}$$

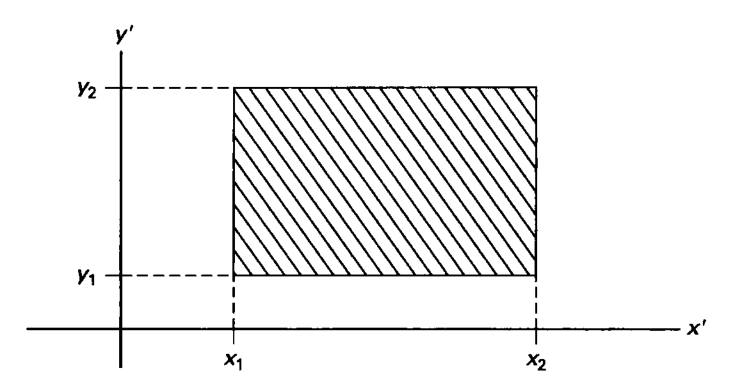


Figure 2.6-5 Point set for the event $\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}$.

According to the elementary properties of the definite integral, the second and third terms on the right-hand side of Equation 2.6-33 can be written, respectively, as

$$\int_{x_1}^{x_2}\int_{-\infty}^{y_1} f_{XY}(\xi,\eta)d\xi\, d\eta = \int_{-\infty}^{x_2}\int_{-\infty}^{y_1} f_{XY}(\xi,\eta)d\xi\, d\eta - \int_{-\infty}^{x_1}\int_{-\infty}^{y_1} f_{XY}(\xi,\eta)d\xi\, d\eta \tag{2.6-34}$$

$$\int_{-\infty}^{x_1}\int_{y_1}^{y_2} f_{XY}(\xi,\eta)d\xi\, d\eta = \int_{-\infty}^{x_1}\int_{-\infty}^{y_2} f_{XY}(\xi,\eta)d\xi\, d\eta - \int_{-\infty}^{x_1}\int_{-\infty}^{y_1} f_{XY}(\xi,\eta)d\xi\, d\eta. \tag{2.6-35}$$

But the terms on the right-hand sides of these equations are all distributions; thus, Equations 2.6-34 and 2.6-35 become

$$\int_{x_1}^{x_2}\int_{-\infty}^{y_1} f_{XY}(\xi,\eta)d\xi\, d\eta = F_{XY}(x_2,y_1) - F_{XY}(x_1,y_1), \tag{2.6-36}$$

$$\int_{-\infty}^{x_1}\int_{y_1}^{y_2} f_{XY}(\xi,\eta)d\xi\, d\eta = F_{XY}(x_1,y_2) - F_{XY}(x_1,y_1). \tag{2.6-37}$$

Now going back to Equation 2.6-33 and using Equations 2.6-36 and 2.6-37 we find that

$$\begin{aligned} F_{XY}(x_2,y_2) = {} & P[x_1 < X \leq x_2, y_1 < Y \leq y_2] \\ & + F_{XY}(x_2,y_1) - F_{XY}(x_1,y_1) + F_{XY}(x_1,y_2) - F_{XY}(x_1,y_1) \\ & + F_{XY}(x_1,y_1). \end{aligned} \tag{2.6-38}$$

After simplifying and rearranging term so that the desired quantity appears on the left-hand side, we finally get

$$\begin{aligned} P[x_1 < X \leq x_2, y_1 < Y \leq y_2] = {} & F_{XY}(x_2,y_2) - F_{XY}(x_2,y_1) \\ & - F_{XY}(x_1,y_2) + F_{XY}(x_1,y_1). \end{aligned} \tag{2.6-39}$$

Equation 2.6-39 is generally true for any random variables X, Y independent or not. Some caution must be taken in applying Equation 2.6-39. For example, Figure 2.6-6(a,b) and (b) show two regions A, B involving excursions on random variables X, Y such that $\{x_1 < X \leq x_2\}$ and $\{y_1 < Y \leq y_2\}$. However, the use of Equation 2.6-39 would not be appropriate here since neither region is a rectangle with sides parallel to the axes. In the case of the event shown in Figure 2.6-6(a), a rotational coordinate transformation might save the day but this would involve some knowledge of transformation of random variables, a subject covered in the next chapter. The events whose point sets are shown in Figure 2.6-6 can still be computed by integration of the probability density function (pdf) provided that the integration is done over the appropriate region. We illustrate with the following example.

Example 2.6-6

(*probabilities for nonrectangular sets*) We are given $f_{XY}(x,y) = e^{-(x+y)}u(x)u(y)$ and wish to compute $P[(X,Y) \in \mathscr{A}]$, where $\mathscr{A}$ is the shaded region shown in Figure 2.6-7. The region $\mathscr{A}$ is described by $\mathscr{A} = \{(x,y)\colon 0 \leq x \leq 1, |y| \leq x\}$. We obtain

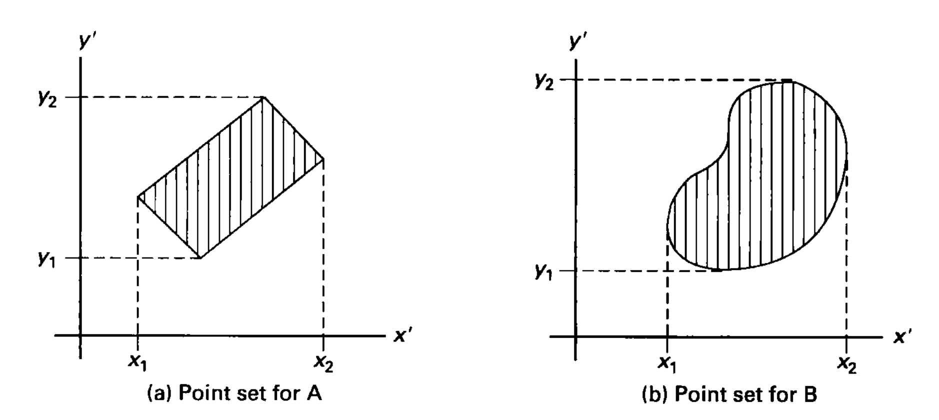


Figure 2.6-6 Points sets of events A and B whose probabilities are not given by Equation 2.6-39.

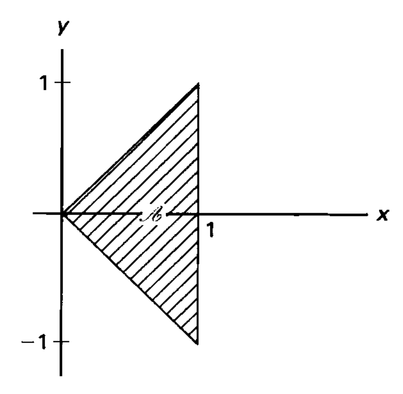
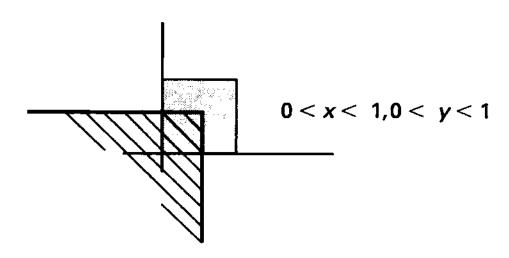


Figure 2.6-7 The region $\mathscr{A}$ for Example 2.6-5.

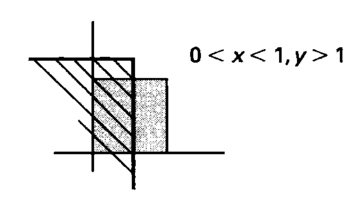
$$
\begin{aligned}
P[(X,Y) \in \mathscr{A}] &= \int_{x=0}^{x=1} \int_{y=-x}^{x} e^{-(x+y)} u(x) u(y) dx\, dy \\
&= \int_0^1 \left(\int_{y=-x}^{x} e^{-y} u(y) dy \right) e^{-x} u(x) dx \\
&= \int_0^1 \left(\int_0^x e^{-y} dy \right) e^{-x} dx \\
&= \int_0^1 (-e^{-y}|_0^x) e^{-x} dx \\
&= \int_0^1 (1 - e^{-x}) e^{-x} dx \\
&= \int_0^1 (e^{-x} - e^{-2x}) dx \\
&= \left(-e^{-x}|_0^1\right) - \left(-\frac{1}{2} e^{-2x}\right)\Big|_0^1 \\
&= 1 - e^{-1} - \frac{1}{2} + \frac{1}{2} e^{-2} \\
&= \frac{1}{2} - e^{-1} + \frac{1}{2} e^{-2} \\
&\doteq 0.1998.
\end{aligned}
\qquad (2.6\text{-}40)
$$

Example 2.6-7

(*computing* CDF) Let X, Y be two random variables with joint pdf $f_{XY}(x,y) = 1$ for $0 < x \leq 1$, $0 < y \leq 1$, and zero elsewhere. The support for the pdf is shown in gray; the support for the event $(-\infty, x] \times (-\infty, y]$ for values $0 < x \leq 1$, $0 < y \leq 1$ is shown bounded by the heavy black line.



For the situation shown in the figure $F_{XY}(x,y) = \int_0^y \int_0^x 1 dx'\, dy' = xy$.
When $0 < x < 1$, $y > 1$, we obtain $F_{XY}(x,y) = \int_0^x dx' \int_0^1 dy = x$. Proceeding in this way, we eventually obtain a complete characterization of the CDF as



$$F_{XY}(x,y) = \begin{cases} 0, & x \leq 0, \text{ or } y \leq 0, \\ xy, & 0 < x \leq 1, 0 < y \leq 1, \\ x, & 0 < x \leq 1, y > 1, \\ y, & x > 1, 0 < y \leq 1, \\ 1, & x > 1, y > 1. \end{cases}$$

As Examples 2.6-6 and 2.6-7 illustrate for specific cases, the probability of any event of the form $\{(X,Y) \in \mathscr{A}\}$ can be computed by the formula

$$P[(X,Y) \in \mathscr{A}] = \iint_{\mathscr{A}} f_{XY}(x,y)dx\,dy \tag{2.6-41}$$

provided $f_{XY}(x,y)$ exists. While Equation 2.6-41 seems entirely reasonable, its veracity requires demonstration. One way to do this is to decompose the arbitrarily shaped region into a (possibly very large) number of tiny disjoint rectangular regions $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_N$. Then the event $\{X, Y \in \mathscr{A}\}$ is decomposed as

$$\{(X,Y) \in \mathscr{A}\} = \bigcup_{i=1}^{N} \{(X,Y) \in \mathscr{A}_i\}$$

with the consequence that (by induced Axiom 3)

$$P[(X,Y) \in \mathscr{A}] = \sum_{i=1}^{N} P[(X,Y) \in \mathscr{A}_i]. \tag{2.6-42}$$

But the probabilities on the right-hand side can be expressed in terms of distributions and hence in terms of integrals of densities (Equation 2.6-39). Then, taking the limit as N becomes large and the $\mathscr{A}_i$ become infinitesimal, we would obtain Equation 2.6-41.

The functions $F_X(x)$ and $F_Y(y)$ are called *marginal* distributions if they are derived from a joint distribution. Thus,

$$F_X(x) = F_{XY}(x,\infty) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{XY}(\xi,y)d\xi dy \tag{2.6-43}$$

$$F_Y(y) = F_{XY}(\infty,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f_{XY}(x,\eta)dxd\eta. \tag{2.6-44}$$

Since the marginal densities are given by

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{2.6-45}$$

$$f_Y(y) = \frac{dF_Y(y)}{dy}, \tag{2.6-46}$$

we obtain the following by partial differentiation of Equation 2.6-43 with respect to x and of Equation 2.6-44 with respect to y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \tag{2.6-47}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx. \tag{2.6-48}$$

We next summarize the key properties of the joint pdf $f_{XY}(x,y)$.

Properties of Joint pdf's.

(i) $f_{XY}(x,y) \geq 0$ for all x, y.

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)dx\,dy = 1$ (the certain event).

(iii) While $f_{XY}(x,y)$ is not a probability, indeed it can be greater than 1, we can regard $f_{XY}(x,y)dx\,dy$ as a differential probability. We will sometimes write $f_{XY}(x,y)\,dx\,dy = P[x < X \leq x + dx, y < Y \leq y + dy]$.

(iv) $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$ and $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx$.

Property (i) follows from the fact that the integral of the joint pdf over any region of the plane must be positive. Also, considering this joint pdf as the mixed partial derivative of the CDF, property (i) easily follows from a limiting operation applied to property (iv) of the joint CDF. Property (ii) follows from the fact that the integral of the joint pdf over the whole plane gives us the probability that the random variables will take on *some value*, which is the certain event with probability 1.

For discrete random variables we obtain similar results. Given the joint PMF $P_{XY}(x_i, y_k)$ for all x_i, y_k, we compute the marginal PMF's from

$$P_X(x_i) = \sum_{\text{all } y_k} P_{XY}(x_i, y_k) \tag{2.6-49}$$

$$P_Y(y_k) = \sum_{\text{all } x_i} P_{XY}(x_i, y_k). \tag{2.6-50}$$

Example 2.6-8

(*waiting time at a restaurant*) A certain restaurant has been found to have the following joint distribution for the waiting time for service for a newly arriving customer and the total

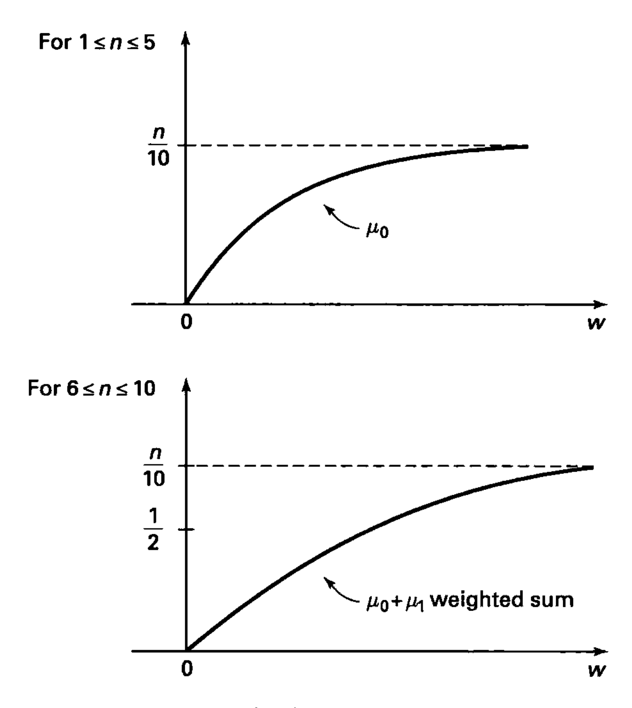


Figure 2.6-8 The CDF of Example 2.6-8: (top) number of customers in the range 1 to 5; (bottom) number of customers in the range 6 to 10.

number of customers including the new arrival. Let W be a random variable representing the continuous waiting time for a newly arriving customer, and let N be a discrete random variable representing the total number of customers.

The joint distribution function is then given as,

$$F_{W,N}(w,n) = \begin{cases} 0, & n < 0 \text{ or } w < 0, \\ (1 - e^{-w/\mu_0})\frac{n}{10}, & 0 \leq n < 5, w \geq 0, \\ (1 - e^{-w/\mu_0})\frac{5}{10} + (1 - e^{-w/\mu_1})\left(\frac{n-5}{10}\right), & 5 \leq n < 10, w \geq 0, \\ (1 - e^{-w/\mu_0})\frac{5}{10} + (1 - e^{-w/\mu_1})\left(\frac{5}{10}\right), & 10 \leq n, w \geq 0 \end{cases},$$

where the parameters μ_i satisfy $0 < \mu_0 < \mu_1$. Note that this choice of the parameters means that waiting times are longer when the number of customers is large.

Noting that W is continuous and N is discrete, we sketch this joint distribution as a function of w for several values of n for $n \geq 0$ and $w \geq 0$ in Figure 2.6-8.

We next find the joint *mixed probability density-mass* function

$$\begin{aligned} f_{W,N}(w,n) &\triangleq \frac{\partial}{\partial w} \nabla_n F_{W,N}(w,n) \\ &= \frac{\partial}{\partial w} \{F_{W,N}(w,n) - F_{W,N}(w,n-1)\} \\ &= \frac{\partial}{\partial w} F_{W,N}(w,n) - \frac{\partial}{\partial w} F_{W,N}(w,n-1). \end{aligned}$$

In words $f_{W,N}(w,n)$ is the pdf of W together or jointly with $\{N = n\}$. Calculating, we obtain

$$\nabla_n F_{W,N}(w,n) = F_{W,N}(w,n) - F_{W,N}(w,n-1) = u(w)\begin{cases}(1-e^{-w/\mu_0})\frac{1}{10}, & 0 < n \leq 5,\\ (1-e^{-w/\mu_1})\frac{1}{10}, & 5 < n \leq 10,\\ 0, & \text{else.}\end{cases}$$

Therefore,

$$\begin{aligned} f_{W,N}(w,n) &= \frac{\partial}{\partial w}\nabla_n F_{W,N}(w,n) \\ &= u(w)\begin{cases}\frac{1}{10}\frac{1}{\mu_0}e^{-w/\mu_0}, & 0 < n \leq 5,\\ \frac{1}{10}\frac{1}{\mu_1}e^{-w/\mu_1}, & 5 < n \leq 10,\\ 0, & \text{else.}\end{cases}\end{aligned}$$

Thus we see a simpler view in terms of the joint pdf, where the shorter average waiting time μ_0 governs the RV W when there are less than or equal to $n = 5$ customers, while the longer average waiting time μ_1 governs when there are more than 5 customers. In a more detailed model, the average waiting time would be expected to increase with each increase in n.

Independent random variables. Two RVs X and Y are said to be *independent* if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for every combination of x, y. In Section 1.5 two events A and B were said to be independent if $P[AB] = P[A]P[B]$. Taking $AB \triangleq \{X \leq x\} \cap \{Y \leq y\}$, where $A \triangleq \{X \leq x\}, B = \{Y \leq y\}$, and recalling that $F_X(x) \triangleq P[X \leq x]$, and so forth for $F_Y(y)$, it then follows immediately that

$$F_{XY}(x,y) = F_X(x)F_Y(y) \tag{2.6-51}$$

for every x, y if and only if X and Y are independent. Also

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x\,\partial y} \tag{2.6-52}$$

$$= \frac{\partial F_X(x)}{\partial x}\cdot\frac{\partial F_Y(y)}{\partial y}$$

$$= f_X(x)f_Y(y). \tag{2.6-53}$$

From the definition of conditional probability we obtain for independent X, Y:

$$F_X(x|Y \leq y) = \frac{F_{XY}(x,y)}{F_Y(y)}$$

$$= F_X(x), \tag{2.6-54}$$

and so forth, for $F_Y(y|X \leq x)$. From these results it follows (by differentiation) that for independent events the conditional pdf's are equal to the marginal pdf's, that is,

$$f_X(x|Y \leq y) = f_X(x) \tag{2.6-55}$$

$$f_Y(y|X \leq x) = f_Y(y). \tag{2.6-56}$$

It is easy to show from Equation 2.6-39 that the events $\{x_1 < X \leq x_2\}$ and $\{y_1 < Y \leq y_2\}$ are independent if X and Y are independent random variables, that is,

$$P[x_1 < X \leq x_2, y_1 < Y \leq y_2] = P[x_1 < X \leq x_2]P[y_1 < Y \leq y_2] \qquad (2.6\text{-}57)$$

if $F_{XY}(x, y) = F_X(x)F_Y(y)$. Indeed, using Equation 2.6-39

$$P[x_1 < X \leq x_2, y_1 < Y \leq y_2]$$
$$= F_{XY}(x_2, y_2) - F_{XY}(x_2, y_1) - F_{XY}(x_1, y_2) + F_{XY}(x_1, y_1) \qquad (2.6\text{-}58)$$
$$= F_X(x_2)F_Y(y_2) - F_X(x_2)F_Y(y_1) - F_X(x_1)F_Y(y_2) + F_X(x_1)F_Y(y_1) \qquad (2.6\text{-}59)$$
$$= (F_X(x_2) - F_X(x_1))(F_Y(y_2) - F_Y(y_1)) \qquad (2.6\text{-}60)$$
$$= P[x_1 < X \leq x_2]P[y_1 < Y \leq y_2]. \qquad (2.6\text{-}61)$$

Example 2.6-9

The experiment consists of throwing a fair die once. The sample space for the experiment is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We define two RVs as follows:

$$X(\zeta) \triangleq \begin{cases} 1 + \zeta, \text{for outcomes } \zeta = 1 \text{ or } 3 \\ 0, \text{ for all other values of } \zeta \end{cases}$$
$$Y(\zeta) \triangleq \begin{cases} 1 - \zeta, \text{for outcomes } \zeta = 1, 2, \text{or } 3 \\ 0, \text{for all other values of } \zeta \end{cases}$$

(a) Compute the relevant single and joint PMFs.
(b) Compute the joint CDFs $F_{XY}(1, 1)$,$F_{XY}(3, -0.5)$, $F_{XY}(5, -1.5)$.
(c) Are the RVs X and Y are independent?

Solution Since the die is assumed fair, each face has a probability of 1/6 of showing up. (a) So the singleton events $\{\zeta\}$ are all equally likely probability $P[\{\zeta\}] = 1/6$. Thus, we obtain

$X(1) = 2, X(3) = 4,$ and for the other outcomes, we have
$X(2) = X(4) = X(5) = X(6) = 0.$

Thus, the PMF P_X is given as $P_X(0) = 4/6, P_X(2) = 1/6, P_X(4) = 1/6$, and $P_X(k) = 0$ for all other k.

Likewise, from the definition of $Y(\zeta)$, we obtain

$$Y(1) = Y(4) = Y(5) = Y(6) = 0,$$

$$Y(2) = -1, \text{and } Y(3) = -2,$$

thus yielding PMF values $P_Y(0) = 4/6, P_Y(-1) = 1/6, P_Y(-2) = 1/6$, and $P_Y(k) = 0$ for all other k.

We next compute the joint PMFs $P_{XY}(i, j)$ directly from the definition, that is, $P_{XY}(i, j) = P[\text{all } \zeta : X(\zeta) = i, Y(\zeta) = j]$. This is easily done if we recall that joint probabilities are probabilities of intersections of subsets of Ω and for example, the event of

observing the die faces of 2, 4, 5, or 6 is written as the subset $\{2, 4, 5, 6\}$. Thus, $P_{XY}(0, 0) = P[\text{all } \zeta : X(\zeta) = 0, Y(\zeta) = 0] = P[\{2, 4, 5, 6\} \cap \{1, 4, 5, 6\}] = P[\{4, 5, 6\}]=1/2$.

Likewise we compute:

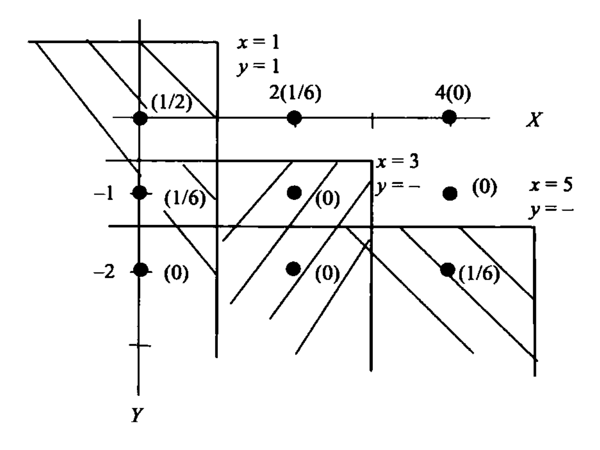
$$\begin{aligned}
P_{XY}(2, 0) &= P[\{1\} \cap \{1, 4, 5, 6\}] = P[\{1\}] = 1/6 \\
P_{XY}(4, 0) &= P[\{3\} \cap \{1, 4, 5, 6\}] = P[\phi] = 0 \\
P_{XY}(0, -1) &= P[\{2, 4, 5, 6\} \cap \{2\}] = P[\{2\}] = 1/6 \\
P_{XY}(2, -1) &= P[\{1\} \cap \{2\}] = P[\phi] = 0 \\
P_{XY}(4, -1) &= P[\{3\} \cap \{2\}] = P[\phi] = 0 \\
P_{XY}(0, -2) &= P[\{2, 4, 5, 6\} \cap \{3\}] = P[\phi] = 0 \\
P_{XY}(2, -2) &= P[\{1\} \cap \{3\}] = P[\phi] = 0 \\
P_{XY}(4, -2) &= P[\{3\} \cap \{3\}] = P[\{3\}] = 1/6.
\end{aligned}$$

(b) For computing the joint CDFs, it is helpful to graph these points and their associated probabilities. These probabilities are shown in parentheses. From the graph we see that

$$F_{XY}(1, 1) = P_{XY}(0, 0) + P_{XY}(0, 1) + P_{XY}(0, 2) = 2/3.$$

Likewise $F_{XY}(3, -0.5) = P_{XY}(0, -1) + P_{XY}(2, -1) + P_{XY}(0, -2) + P_{XY}(2, -2) = \frac{1}{6}$ and $F_{XY}(5, -1.5) = P_{XY}(0, -2) + P_{XY}(2, -2) + P_{XY}(4, -2) = \frac{1}{6}$.

(c) To check for dependence, it is sufficient to find one point where the pdf (or CDF) does not factor. Consider then $P_{XY}(2, 0) = 1/6$, but $P_X(2)P_Y(0) = 1/6 \times 4/6 = 1/9$, so the random variables X and Y are not independent.



Probabilities associated with the Example 2.6-9.

Example 2.6-10

(joint pdf of independent Gaussians)

$$\begin{aligned}
f_{XY}(x, y) &= \frac{1}{2\pi\sigma^2} e^{-(1/2\sigma^2)(x^2+y^2)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x^2/\sigma^2)} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(y^2/\sigma^2)}.
\end{aligned} \tag{2.6-62}$$

Hence X and Y are independent RVs.

Probabilities associated with the example 2.6-9. The numbers in parenthesis are the probabilities of reaching those points. For example, $P_{XY}(0,0) = 1/2$.

Example 2.6-11

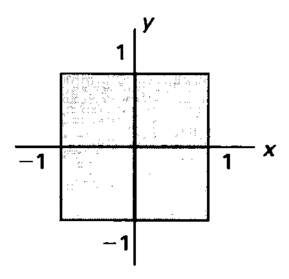
(*calculations with independent Gaussians*) The joint pdf of two random variables is given by $f_{XY}(x,y) = [2\pi]^{-1}\exp[-\frac{1}{2}(x^2+y^2)]$ for $-\infty < x, y < \infty$. Compute the probability that both X and Y are restricted to (a) the 2×2 square; and (b) the unit circle.

Solution (a) Let $\Re_1$ denote the surface of the square. Then

$$P[\zeta\colon (X,Y) \in \Re_1] = \iint_{\Re_1} f_{XY}(x,y)dxdy \tag{2.6-63}$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-1}^{1}\exp\left[-\frac{1}{2}x^2\right]dx \times \frac{1}{\sqrt{2\pi}}\int_{-1}^{1}\exp\left[-\frac{1}{2}y^2\right]dy \tag{2.6-64}$$

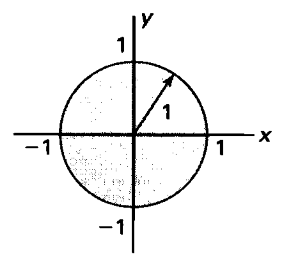
$$= 2\text{erf}(1) \times 2\text{erf}(1) = 0.465. \tag{2.6-65}$$



(b) Let $\Re_2$ denote the surface of the unit circle. Then

$$P[\zeta\colon (X,Y) \in \Re_2] = \iint_{\Re_2} f_{XY}(x,y)dxdy \tag{2.6-66}$$

$$= \iint_{\Re_2} [2\pi]^{-1}\exp\left[-\frac{1}{2}(x^2+y^2)\right]dxdy. \tag{2.6-67}$$



With the substitution $r \triangleq \sqrt{x^2+y^2}$ and $\tan\theta \triangleq y/x$, the infinitesimal area $dxdy \to rdrd\theta$, and we obtain

$$P[\zeta: (X,Y) \in \Re_2] = \frac{1}{2\pi} \iint_{\Re_2} \exp\left(-\frac{1}{2}r^2\right) r\, dr\, d\theta \tag{2.6-68}$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \left[\int_0^1 r \exp\left(-\frac{1}{2}r^2\right) dr\right] d\theta \tag{2.6-69}$$

$$= \int_0^1 r \exp\left(-\frac{1}{2}r^2\right) dr \tag{2.6-70}$$

$$= \int_0^{1/2} e^{-z} dz, \qquad \text{with} \quad z \triangleq \frac{1}{2}r^2, \tag{2.6-71}$$

$$\doteq 0.393. \tag{2.6-72}$$

Joint densities involving nonindependent RVs. Lest the reader think that all joint CDF's or pdf's factor, we next consider a case involving nonindependent random variables.

Example 2.6-12

(*computing joint* CDF) Consider the simple but nonfactorable joint pdf

$$f_{XY}(x,y) = A(x+y) \qquad 0 < x \leq 1, \qquad 0 < y \leq 1, \tag{2.6-73}$$

$$= 0, \qquad \text{otherwise,} \tag{2.6-74}$$

and answer the following questions.

(i) What is A? We know that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) dx\, dy = 1.$$

Hence

$$A \int_0^1 dy \int_0^1 x\, dx + A \int_0^1 dx \int_0^1 y\, dy = 1 \Rightarrow A = 1.$$

(ii) What are the marginal pdf's?

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y) dy = \int_0^1 (x+y) dy = (xy + y^2/2)\Big|_0^1$$

$$= \begin{cases} x + \frac{1}{2}, & 0 < x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y) dx$$

$$= \begin{cases} y + \frac{1}{2}, & 0 < y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

(iii) What is $F_{XY}(x, y)$? $F_{XY}(x, y) \triangleq P[X \leq x, Y \leq y]$, so we must integrate over the infinite rectangle with vertices (x, y), $(x, -\infty)$, $(-\infty, -\infty)$, and $(-\infty, y)$. However, only where this rectangle actually overlaps with the region over which $f_{XY}(x, y) \neq 0$, that is, the support of the pdf written supp(f) will there be a contribution to the integral

$$F_{XY}(x, y) = \int_{-\infty}^{x} dx' \int_{-\infty}^{y} dy' f_{XY}(x', y').$$

(a) $x \geq 1$, $y \geq 1$ [Figure 2.6-9(a)]

$$F_{XY}(x, y) = \int_0^1 \int_0^1 f_{XY}(x', y') dx'\, dy' = 1.$$

(b) $0 < x \leq 1$, $y \geq 1$ [Figure 2.6-9(b)]

$$\begin{aligned} F_{XY}(x, y) &= \int_{y'=0}^{1} dy' \left(\int_{x'=0}^{x} dx'(x' + y') \right) \\ &= \int_{y'=0}^{1} dy' \left(\int_{x'=0}^{x} x'\, dx \right) + \int_{y'=0}^{1} dy'\, y' \left(\int_{x'=0}^{x} dx' \right) \\ &= \frac{x}{2}(x + 1). \end{aligned}$$

(c) $0 < y \leq 1$, $x \geq 1$ [Figure 2.6-9(c)]

$$F_{XY}(x, y) = \int_{y'=0}^{y} \int_{x'=0}^{1} (x' + y') dx'\, dy' = \frac{y}{2}(y + 1).$$

(d) $0 < x \leq 1$, $0 < y \leq 1$ [Figure 2.6-9(d)]

$$F_{XY}(x, y) = \int_{y'=0}^{y} \int_{x'=0}^{x} (x' + y') dx'\, dy' = \frac{yx}{2}(x + y).$$

(e) $x < 0$, for any y; or $y < 0$, for any x [Figure 2.6-9(e)]

$$F_{XY}(x, y) = 0.$$

(f) Compute $P[X + Y \leq 1]$. The point set is the half-space separated by the line $x + y = 1$ or $y = 1 - x$. However, only where this half-space intersects the region over which $f_{XY}(x, y) \neq 0$, will there be a contribution to the integral

$$P[X + Y \leq 1] = \iint_{x'+y' \leq 1} f_{XY}(x', y') dx'\, dy'.$$

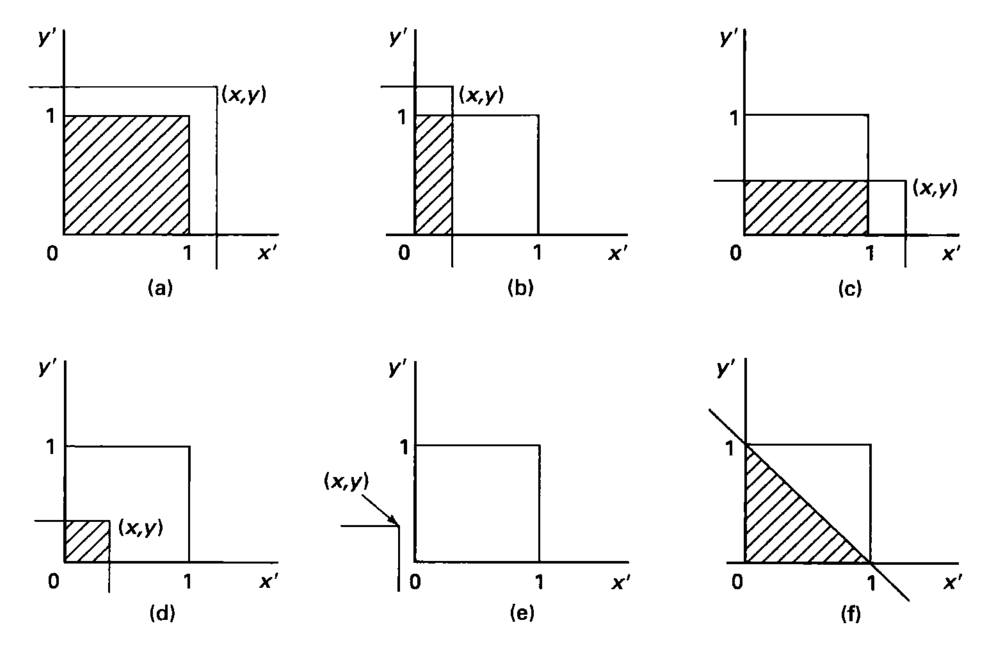


Figure 2.6-9 Shaded region in (a) to (e) is the intersection of supp(f_{XY}) with the point set associated with the event $\{-\infty < X \le x, -\infty < Y \le y\}$. In (f), the shaded region is the intersection of supp(f_{XY}) with $\{X+Y \le 1\}$.

[See Figure 2.6-9(f).] Hence

$$\begin{aligned} P[X+Y \le 1] &= \int_{x'=0}^{1} \int_{y'=0}^{1-x'} (x'+y')dy'\,dx' \\ &= \int_{x'=0}^{1} x'(1-x')dx' + \int_{x'=0}^{1} \frac{(1-x')^2}{2} dx' \\ &= \frac{1}{3}. \end{aligned}$$

In the previous example we dealt with a pdf that was not factorable. Another example of a joint pdf that is not factorable is

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2\sigma^2(1-\rho^2)}(x^2+y^2-2\rho xy)\right) \tag{2.6-75}$$

when $\rho \neq 0$. In this case, X and Y are not independent.

In the special case when $\rho = 0$ in Equation 2.6-75, $f_{XY}(x,y)$ factors as $f_X(x)f_Y(y)$ and X and Y become independent random variables. A picture of $f_{XY}(x,y)$ under these circumstances is shown in Figure 2.6-10 for $\sigma = 1$.

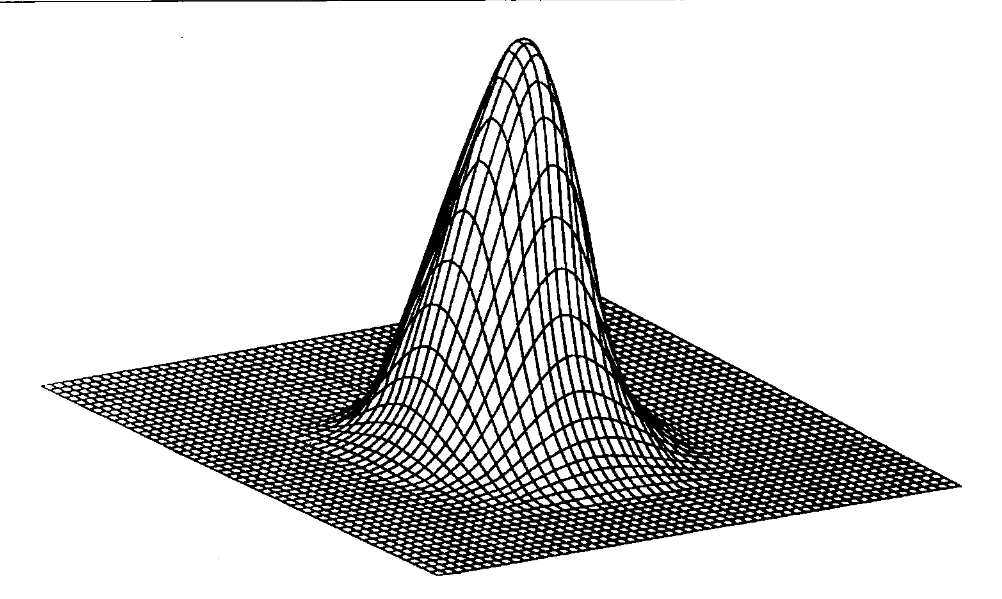


Figure 2.6-10 Graph of the joint Gaussian density

$$f_{XY}(x,y) = (2\pi)^{-1} \exp\left[-\frac{1}{2}(x^2+y^2)\right].$$

As we shall see in Chapter 4, Equation 2.6-75 is a special case of the *jointly Gaussian probability density* of two RVs. We defer a fuller discussion of this important pdf until we discuss the meaning of the parameter ρ. This we do in Chapter 4. We shall see in Chapter 5 that Equation 2.6-75 and its generalization can be written compactly in matrix form.

Example 2.6-13

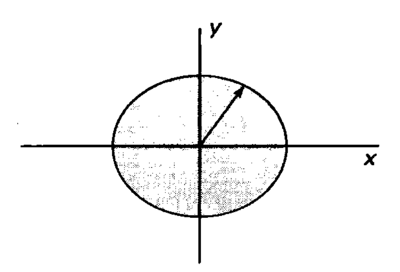
(*calculation with dependent Gaussian RVs*) Consider again the problem considered in Example 2.6-11, part (b), except let

$$f_{XY}(x,y) = [2\pi\sqrt{1-\rho^2}]^{-1} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right].$$

As before let $\Re_2$ denote the surface of the unit circle. Then

$$P[\zeta\colon (X,Y) \in \Re_2] = \iint_{\Re_2} f_{XY}(x,y)dxdy \tag{2.6-76}$$

$$= \iint_{\Re_2} [2\pi\sqrt{1-\rho^2}]^{-1} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right) dx\, dy. \tag{2.6-77}$$



With the polar coordinate substitution $r \triangleq \sqrt{x^2+y^2}$, $\tan\theta \triangleq y/x$, we obtain

$$\begin{aligned}
P[\zeta\colon (X,Y) \in \Re_2] &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^{2\pi}\int_0^1 r \exp\left(-\frac{1}{2(1-\rho^2)}(r^2 - 2\rho r^2 \cos\theta \sin\theta)\right) dr d\theta \\
&= \frac{K}{\pi}\int_0^{2\pi}\int_0^1 r\exp\left(-r^2\left[2K^2(1-\rho\sin 2\theta)\right]\right) dr d\theta \quad \text{with} \\
& K \triangleq \frac{1}{2\sqrt{1-\rho^2}} \quad \text{and} \quad \sin 2\theta = 2\sin\theta\cos\theta, \\
&= \frac{K}{2\pi}\int_0^{2\pi}\left[\int_0^1 \exp -\left(\left[2K^2(1-\rho\sin 2\theta)\right] z\right) dz\right] d\theta \text{ with } \quad z \triangleq r^2, \\
&= \frac{K}{2\pi}\int_0^{2\pi} \frac{1}{2K^2(1-\rho\sin 2\theta)}(1 - \exp -\left[2K^2(1-\rho\sin 2\theta)\right] d\theta \\
&= \frac{1}{4\pi K}\int_0^{2\pi}\left(\frac{1-\exp[-2K^2(1-\rho\sin 2\theta)]}{1-\rho\sin 2\theta}\right) d\theta.
\end{aligned}$$

For $\rho = 0$, we get the probability 0.393, that is, the same as in Example 2.6-10. However, when $\rho \neq 0$, this probability must be computed numerically since this integral is not available in closed form. A MATLAB.m file that enables the computation of $P(\zeta\colon (X,Y) \in \Re_2)$ is furnished below. The result is shown in Figure 2.6-11.

MATLAB.m file for computing. $P[\zeta\colon (X,Y) \in \Re_2]$

```
function[Pr]=corrprob
p=[0:100]/100.;
q=p*2*pi;
Pr=zeros(1,100);

K=.5./sqrt(1-p.^2);

for i=1:100
   f=(1-exp(-2*K(i)^2*(1-p(i)*sin(2*q)))).(1-p(i)*sin(2*q));
   Pr(i)=sum(f)/(4*pi)/K(i)*(2*pi/100);
end
```

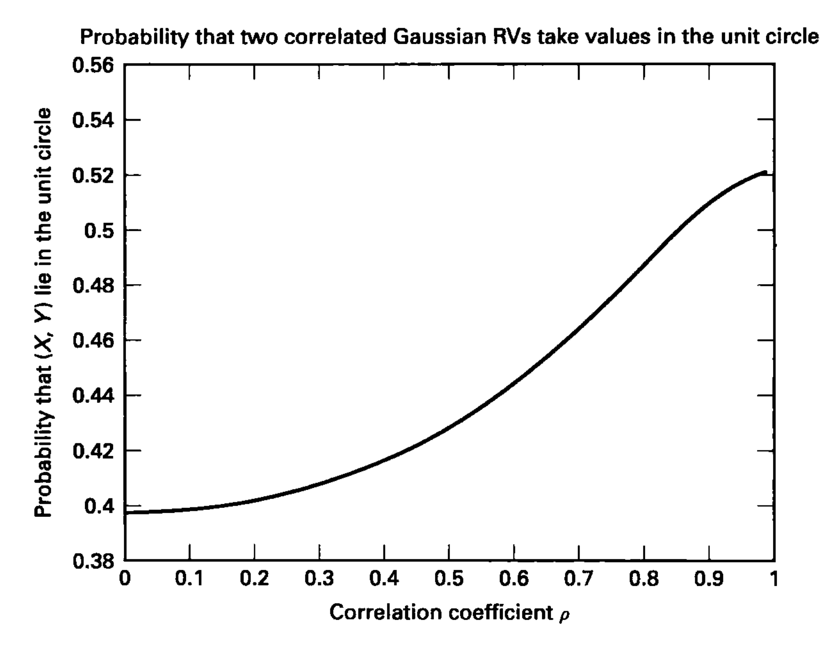


Figure 2.6-11 Result of MATLAB computation in Example 2.6-13.

```
plot(p(1:100),Pr)
title('Probability that two correlated Gaussian RVs take values in the
  unit circle')
xlabel('Correlation coefficient rho')
ylabel('Probability that X,Y) lie in the unit circle')
```

In Section 4.3 of Chapter 4 we demonstrate the fact that as $\rho \to 1$
$f_{XY}(x,y) \to \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\delta(y-x)$. Hence

$$P[\zeta : (X,Y) \in \Re_2] = \iint_{\Re_2} \frac{1}{\sqrt{2\pi}} \tag{2.6-78}$$

$$e^{-0.5x^2}\delta(x-y)dxdy = \int_{-0.707}^{0.707} \frac{1}{\sqrt{2\pi}}e^{-0.5x^2}dx \tag{2.6-79}$$

$$= 0.52. \tag{2.6-80}$$

This is the result that we observe in Figure 2.6-11.

Conditional densities. We shall now derive a useful formula for conditional densities involving two RVs. The formula is based on the definition of conditional probability given in Equation 1.6-2. From Equation 2.6-39 we obtain

$$P[x < X \le x + \Delta x, y < Y \le y + \Delta y]$$
$$= F_{XY}(x + \Delta x, y + \Delta y) - F_{XY}(x, y + \Delta y) - F_{XY}(x + \Delta x, y) + F_{XY}(x, y). \quad (2.6\text{-}81)$$

Now dividing both sides by $\Delta x\,\Delta y$, taking limits, and subsequently recognizing that the right-hand side, by definition, is the second partial derivative of F_{XY} with respect to x and y enables us to write that

$$\lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{P[x < X \le x + \Delta x, y < Y \le y + \Delta y]}{\Delta x\,\Delta y} = \frac{\partial^2 F_{XY}}{\partial x\,\partial y} \triangleq f_{XY}(x, y).$$

Hence for Δx, Δy small

$$P[x < X \le x + \Delta x, y < Y \le y + \Delta y] \simeq f_{XY}(x, y)\Delta x\,\Delta y, \quad (2.6\text{-}82)$$

which is the two-dimensional equivalent of Equation 2.4-6. Now consider

$$P[y < Y \le y + \Delta y | x < X \le x + \Delta x] = \frac{P[x < X \le x + \Delta x, y < Y \le y + \Delta y]}{P[x < X \le x + \Delta x]}$$
$$\simeq \frac{f_{XY}(x, y)\Delta x\,\Delta y}{f_X(x)\Delta x}. \quad (2.6\text{-}83)$$

But the quantity on the left is merely

$$F_{Y|B}(y + \Delta y | x < X \le x + \Delta x) - F_{Y|B}(y | x < X \le x + \Delta x),$$

where $B \triangleq \{x < X \le x + \Delta x\}$. Hence

$$\lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{F_{Y|B}(y + \Delta y | x < X \le x + \Delta x) - F_{Y|B}(y | x < X \le x + \Delta x)}{\Delta y}$$
$$= \frac{f_{XY}(x, y)}{f_X(x)}$$
$$= \frac{\partial F_{X|Y}(y | X = x)}{\partial y}$$
$$= f_{Y|X}(y|x) \quad (2.6\text{-}84)$$

by Equation 2.6-3. The notation $f_{Y|X}(y|x)$ reminds us that it is the conditional pdf of Y given the event $\{X = x\}$. We thus obtain the important formula

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \qquad f_X(x) \ne 0. \quad (2.6\text{-}85)$$

If we use Equation 2.6-85 in Equation 2.6-48 we obtain the useful formula:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx. \quad (2.6\text{-}86)$$

Also

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}, \qquad f_Y(y) \neq 0. \tag{2.6-87}$$

The quantity $f_{X|Y}(x|y)$ is called the conditional pdf of X given the event $\{Y = y\}$. From Equations 2.6-85 and 2.6-86 it follows that

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}. \tag{2.6-88}$$

We illustrate with an example.

Example 2.6-14

(*laser coherence*) Suppose we observe the light field $U(t)$ being emitted from a laser. Laser light is said to be *temporally coherent*, which means that the light at any two times t_1 and t_2 is statistically dependent if $t_2 - t_1$ is not too large [2-5]. Let $X \triangleq U(t_1)$, $Y \triangleq U(t_2)$ and $t_2 > t_1$. Suppose X and Y are modeled as jointly Gaussian† as in Equation 2.6-75 with $\sigma^2 = 1$. For $\rho \neq 0$, they are dependent, it turns out that using the defining Equations 2.6-47 and 2.6-48, one can show the marginal densities $f_X(x)$ and $f_Y(y)$ are individually Gaussian. We defer the proof of this to Chapter 4. Since the means are both zero here and the variances are both one, we get for the marginal densities

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \quad \text{and} \quad f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2},$$

both are centered about zero. Now suppose that we measure the light at t_1, that is, X and find that $X = x$. Is the pdf of Y, conditioned upon this new knowledge, still centered at zero, that is, is the average‡ value of Y still zero?

Solution We wish to compute $f_{Y|X}(y|x)$.

Applying Equation 2.6-85,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

yields

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{\left[-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right] + \frac{1}{2}x^2\right\}.$$

†Light is often modeled by Poisson distribution due to its photon nature. As seen in Chapter 1, for a large photon count, the Gaussian distribution well approximates the Poisson distribution. Of course light intensity cannot be negative, but if the mean is large compared to the standard deviation ($\mu >> \sigma$), then the Gaussian density will be very small there.

‡A concept to be fully developed in Chapter 4.

If we multiply and divide the isolated $\frac{1}{2}x^2$ term in the far right of the exponent by $1-\rho^2$, we simplify the above as

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\left[\frac{x^2+y^2-2\rho xy - x^2(1-\rho^2)}{2(1-\rho^2)}\right]\right\}.$$

Further simplifications result when quadratic terms in the exponent are combined into a perfect square:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right).$$

Thus, when $X = x$, the pdf of Y is centered at $y = \rho x$ and not zero as previously. If $\rho x > 0$, Y is more likely to take on positive values, and if $\rho x < 0$, Y is more likely to take on negative values. This is in contrast to what happens when X is not observed: The most likely value of Y is then zero!

A major application of conditioned events and conditional probabilities occurs in the science of estimating failure rates. This is discussed in the next section.

2.7 FAILURE RATES

In modern industrialized society where planning for equipment replacement, issuance of life insurance, and so on are important activities, there is a need to keep careful records of the failure rates of objects, be they machines or humans. For example consider the cost of life insurance: Clearly it wouldn't make much economic sense to price a five-year term-life insurance policy for a 25-year-old woman at the same level as, say, for a 75-year-old man. The "failure" probability (i.e., death) for the older man is much higher than for the young woman. Hence, sound pricing policy will require the insurance company to insure the older man at a higher price. How much higher? This is determined from actuarial tables which are estimates of life expectancy conditioned on many factors. One important condition is "that you have survived until (a certain age)." In other words, the probability that you will survive to age 86, given that you have survived to age 85, is much higher than the probability that you will survive to age 86 if you are an infant.

Let X denote the time of failure or, equivalently, the failure time. Then by Bayes' theorem, the probability that failure will occur in the interval $[t, t+dt]$ *given that* the object has survived to t can be written as

$$P[t < X \leq t+dt | X > t] = \frac{P[t < X \leq t+dt, X > t]}{P[X > t]}. \tag{2.7-1}$$

But since the event $\{X > t\}$ is subsumed by the event $\{t < X \leq t+dt\}$, it follows that $P[t < X \leq t+dt, X > t] = P[t < X \leq t+dt]$. Hence

$$P[t < X \leq t+dt | X > t] = \frac{P[t < X \leq t+dt]}{P[X > t]}. \tag{2.7-2}$$

By recalling that $P[t < X \leq t + dt] = F_X(t + dt) - F_X(t)$, we obtain

$$P[t < X \leq t + dt | X > t] = \frac{F_X(t + dt) - F_X(t)}{1 - F_X(t)}. \tag{2.7-3}$$

A Taylor series expansion of the CDF $F_X(t + dt)$ about the point t yields (we assume that F_X is differentiable)

$$F_X(t + dt) = F_X(t) + f_X(t)\, dt.$$

When this result is used in Equation 2.7-3, we obtain at last

$$\begin{aligned} P[t < X \leq t + dt | X > t] &= \frac{f_X(t)dt}{1 - F_X(t)} \\ &\triangleq \alpha(t)\, dt, \end{aligned} \tag{2.7-4}$$

where

$$\alpha(t) \triangleq \frac{f_X(t)}{1 - F_X(t)}. \tag{2.7-5}$$

The object $\alpha(t)$ is called the *conditional failure rate* although it has other names such as the hazard rate, force of mortality, intensity rate, instantaneous failure rate, or simply failure rate. If the conditional failure rate at t is large, then an object surviving to time t will have a higher probability of failure in the next Δt seconds than another object with lower conditional failure rate. Many objects, including humans, have failure rates that vary with time. During the early life of the object, failure rates may be high due to inherent or congenital defects. After this early period, the object enjoys a useful life characterized by a near-constant failure rate. Finally, as the object ages and parts wear out, the failure rate increases sharply leading quickly and inexorably to failure or death.

The pdf of the random variable X can be computed explicitly from Equation 2.7-3 when we observe that $F_X(t + dt) - F_X(t) = F'_X(t)dt = dF_X$. Thus, we get

$$\frac{dF_X}{1 - F_X} = \alpha(t)\, dt, \tag{2.7-6}$$

which can be solved by integration. First recall from calculus that

$$\int_{y_0}^{y_1} \frac{dy}{1 - y} = -\int_{1-y_0}^{1-y_1} \frac{dy}{y} = \int_{1-y_1}^{1-y_0} \frac{dy}{y} = \ln \frac{1 - y_0}{1 - y_1}.$$

Second, use the facts that

(i) $F_X(0) = 0$ since we assume that the object is working at $t = 0$ (the time that the object is turned on);

(ii) $F_X(\infty) = 1$ since we assume that the object must ultimately fail. Then

$$\int_{F_X(0)}^{F_X(t)} \frac{dF_X}{1 - F_X} = -\ln[1 - F_X(t)] = \int_0^t \alpha(t')dt',$$

from which we finally obtain

$$F_X(t) = 1 - \exp\left[-\int_0^t \alpha(t')dt'\right]. \tag{2.7-7}$$

Since $F_X(\infty) = 1$ we must have

$$\int_0^\infty \alpha(t')\,dt' = \infty. \tag{2.7-8}$$

Equation 2.7-7 is the CDF for the failure time X. By differentiating Equation 2.7-7, we obtain the pdf

$$f_X(t) = \alpha(t)\exp\left[-\int_0^t \alpha(t')dt'\right]. \tag{2.7-9}$$

Different pdf's result from different models for the conditional failure rate $\alpha(t)$.

Example 2.7-1

(*conditional failure rate for the exponential case*) Assume that X obeys the exponential probability law, that is, $F_X(t) = (1 - e^{-\lambda t})u(t)$. We find

$$\alpha(t) = \frac{f_X(t)}{1 - F_X(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda.$$

Thus, the conditional failure rate is a constant. Conversely, if $\alpha(t)$ is a constant, the failure time obeys the exponential probability law.

An important point to observe is that the conditional failure rate is not a pdf (see Equation 2.7-8). The conditional density of X, given $\{X \geq t\}$, can be computed from the conditional distribution by differentiation. For example,

$$\begin{aligned} F_X(x|X > t) &\triangleq P[X \leq x|X > t] \\ &= \frac{P[X \leq x, X > t]}{P[X > t]}. \end{aligned} \tag{2.7-10}$$

The event $\{X \leq x, X > t\}$ is clearly empty if $t > x$. If $t < x$, then $\{X \leq x, X > t\} = \{t < X \leq x\}$. Thus,

$$F_X(x|X > t) = \begin{cases} 0, & t > x, \\ \dfrac{F_X(x) - F_X(t)}{1 - F_X(t)}, & x \geq t. \end{cases} \tag{2.7-11}$$

Hence

$$f_X(x|X \geq t) = \begin{cases} 0, & t > x, \\ \dfrac{f_X(x)}{1 - F_X(t)}, & x \geq t. \end{cases} \tag{2.7-12}$$

The connection between $\alpha(t)$ and $f_X(x|X \geq t)$ is obtained by comparing Equation 2.7-12 with Equation 2.7-5, that is,

$$f_X(t|X \geq t) = \alpha(t). \tag{2.7-13}$$

Example 2.7-2

(*Itsibitsi breakdown*) Oscar, a college student, has a nine-year-old Itsibitsi, an import car famous for its reliability. The conditional failure rate, based on field data, is $\alpha(t) = 0.06tu(t)$ assuming a normal usage of 10,000 mile/year. To celebrate the end of the school year, Oscar begins a 30-day cross-country motor trip. What is the probability that Oscar's Itsibitsi will have a first breakdown during his trip?

Solution First, we compute the pdf $f_X(t)$ as

$$f_X(t) = 0.06te^{-\int_0^t 0.06t'\,dt'}u(t) \tag{2.7-14}$$

$$= 0.06te^{-0.03t^2}u(t). \tag{2.7-15}$$

Next, we convert 30 days into 0.0824 years. Finally, we note that

$$P[9.0 < X \leq 9.0824|X > 9] = \frac{P[9.0 < X \leq 9.0824]}{1 - F_X(9)},$$

where we have used Bayes' rule and the fact the event $\{9 < X \leq 9.0824\} \cap \{X > 9\} = \{9 < X \leq 9.0824\}$.

Since

$$P[9.0 < X \leq 9.0824] = 0.06 \int_{9.0}^{9.0824} te^{-0.03t^2}\,dt \tag{2.7-16}$$

$$= \frac{0.06}{2} \int_{(9.0)^2}^{(9.0824)^2} e^{-0.03z}\,dz \quad \text{with} \quad z \triangleq t^2, \tag{2.7-17}$$

$$= \frac{0.06}{2}\frac{1}{0.03}\left(e^{-0.03(9.0)^2} - e^{-0.03(9.0824)^2}\right) \tag{2.7-18}$$

$$= \left(e^{-0.03(9.0)^2} - e^{-0.03(9.0824)^2}\right) \tag{2.7-19}$$

$$\simeq 0.0038 \tag{2.7-20}$$

and

$$1 - F_X(9) = 0.088,$$

Oscar's car has a $3.8 \times 10^{-3}/8.8 \times 10^{-2}$ or 0.043 probability of suffering a breakdown in the next 30 days.

Incidentally, the probability that a newly purchased Itsibitsi will have at least one breakdown in ten years is 0.95.

SUMMARY

The material discussed in this chapter is central to the concept of the whole book. We began by defining a real random variable as a mapping from the sample space Ω to the real line R. We then introduced a point function $F_X(x)$ called the cumulative distribution function (CDF), which enabled us to compute the probabilities of events of the type $\{\zeta\colon \zeta \in \Omega, X(\zeta) \leq x\}$. The probability density function (pdf) and probability mass function (PMF) were derived from the CDF, and a number of useful and specific probability laws were discussed. We showed how, by using Dirac delta functions, we could develop a unified theory for both discrete and continuous random variables. We then discussed joint distributions, the Poisson transform, and its inverse and the application of these concepts to physical problems.

We discussed the important concept of conditional probability and illustrated its application in the area of conditional failure rates. The conditional failure, often high at the outset, constant during mid-life, and high at old age, is fundamental in determining the probability law of time-to-failure.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

2.1 The event of k successes in n tries regardless of the order is the binomial law $b(k, n; p)$. Let $n = 10$, $p = 0.4$. Define the RV X by

$$X(k) = \begin{cases} 1, & \text{for } 0 \leq k \leq 2, \\ 2, & \text{for } 2 < k \leq 5, \\ 3, & \text{for } 5 < k \leq 8, \\ 4, & \text{for } 8 < k \leq 10. \end{cases}$$

Compute the probabilities $P[X = j]$ for $j = 1, \ldots, 4$. Plot the CDF $F_X(x) = P[X \leq x]$ for all x.

***2.2** Consider the probability space $(\Omega, \mathcal{F}, P)$. Give an example, and substantiate it in a sentence or two, where all outcomes have probability zero. Hint: Think in terms of random variables.

2.3 In a restaurant known for its unusual service, the time X, in minutes, that a customer has to wait before he captures the attention of a waiter is specified by the following CDF:

$$F_X(x) = \begin{cases} \left(\frac{x}{2}\right)^2, & \text{for } 0 \le x \le 1, \\ \frac{x}{4}, & \text{for } 1 \le x \le 2, \\ \frac{1}{2}, & \text{for } 2 \le x \le 10, \\ \frac{x}{20}, & \text{for } 10 \le x \le 20, \\ 1, & \text{for } x \ge 20. \end{cases}$$

(a) Sketch $F_X(x)$. (b) Compute and sketch the pdf $f_X(x)$. Verify that the area under the pdf is indeed unity. (c) What is the probability that the customer will have to wait (1) at least 15 minutes, (2) less than 5 minutes, (3) between 5 and 15 minutes, (4) exactly 1 minute?

2.4 Compute the probabilities of the events $\{X < a\}$, $\{X \le a\}$, $\{a \le X < b\}$, $\{a \le X \le b\}$, $\{a < X \le b\}$, and $\{a < X < b\}$ in terms of $F_X(x)$ and $P[X = x]$ for $x = a, b$.

2.5 In the following pdf's, compute the constant B required for proper normalization:
Cauchy ($\alpha < \infty, \beta > 0$):

$$f_X(x) = \frac{B}{1 + [(x - \alpha)/\beta]^2}, \qquad -\infty < x < \infty.$$

Maxwell ($\alpha > 0$):

$$f_X(x) = \begin{cases} Bx^2 e^{-x^2/\alpha^2}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

***2.6** For these more advanced pdf's, compute the constant B required for proper normalization:
Beta ($b > -1, c > -1$):

$$f_X(x) = \begin{cases} Bx^b(1 - x)^c, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

(See formula 6.2-1 on page 258 of [2-6].)
Chi-square ($\sigma > 0, n = 1, 2, \ldots$):

$$f_X(x) = \begin{cases} Bx^{(n/2)-1} \exp(-x/2\sigma^2), & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

2.7 Let X be a continuous random variable with pdf

$$f_X(x) = \begin{cases} kx, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

where k is a constant.

(a) Determine the value of k and sketch $f_X(x)$.
(b) Find and sketch the corresponding CDF $F_X(x)$.
(c) Find P ($1/4 < X \le 2$)

2.8 Compute $F_X(k\sigma)$ for the Rayleigh pdf (Equation 2.4-15) for $k = 0, 1, 2, \ldots$.

2.9 Write the *probability density functions* (using delta functions) for the Bernoulli, binomial, and Poisson PMF's.

2.10 The pdf of a RV X is shown in Figure P2.10. The numbers in parentheses indicate area. (a) Compute the value of A; (b) sketch the CDF; (c) compute $P[2 \leq X < 3]$;

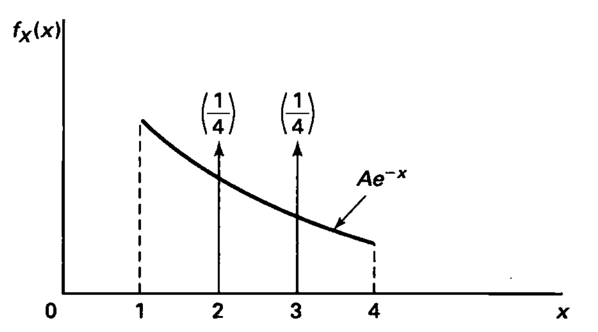


Figure P2.10 pdf of a Mixed RV.

(d) compute $P[2 < X \leq 3]$; (e) compute $F_X(3)$.

2.11 The CDF of a random variable X is given by $F_X(x) = (1 - e^{-4x})u(x)$. Find the probability of the event $\{\zeta \colon X(\zeta) < 1 \text{ or } X(\zeta) > 2\}$.

2.12 The pdf of random variable X is shown in Figure P2.10. The numbers in parentheses indicate area. Compute the value of A. Compute $P[2 < X < 4]$.

2.13 (*two coins tossing*) The experiment consists of throwing two indistinguishable coins simultaneously. The sample space is $\Omega =$\{two heads, one head, no heads\}, which we denote abstractly as $\Omega = \{\zeta_1, \zeta_2, \zeta_3\}$. Next, define two random variables as

$$X(\zeta_1) = 0,\ X(\zeta_2) = 0,\ X(\zeta_3) = 1$$
$$Y(\zeta_1) = 1,\ Y(\zeta_2) = -1,\ Y(\zeta_3) = 1.$$

(a) Compute all possible joint probabilities of the form $P[\zeta : X(\zeta) = \alpha, Y(\zeta) = \beta]$, $\alpha \in \{0, 1\}$, $\beta \in \{-1, 1\}$.

(b) Determine whether X and Y are independent random variables.

2.14 The pdf of the random variable X is shown in Figure P2.14. The numbers in parentheses indicate the corresponding impulse area.
So,

$$f_X(x) = \frac{1}{8}\delta(x+2) + \frac{1}{16}\delta(x+1) + \frac{1}{16}\delta(x-1) + \begin{cases} Ax^2, & |x| \leq 2, \\ 0, & \text{else.} \end{cases}$$

Note that the density f_X is zero off of $[-2, +2]$.

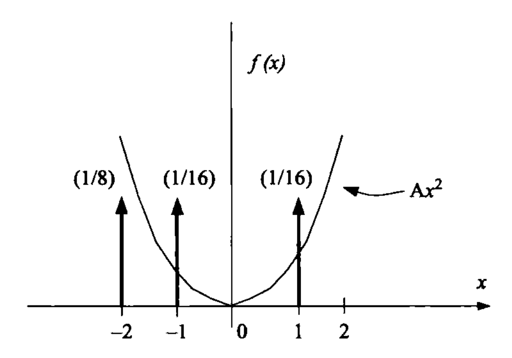


Figure P2.14 pdf of the Mixed pv in the problem 2.14.

(a) Determine the value of the constant A.
(b) Plot the CDF $F_X(x)$. Please label the significant points on your plot.
(c) Calculate $F_X(1)$.
(d) Find $P[-1 < X \leq 2]$.

2.15 Consider a binomial RV with PMF $b(k; 4, \frac{1}{2})$. Compute $P[X = k|X \text{ odd}]$ for $k = 0, \ldots, 4$.

***2.16** Continuing with Example 2.6-8, find the marginal distribution function $F_N(n)$. Find and sketch the corresponding PMF $P_N(n)$. Also find the conditional probability density function $f_W(w|N = n) = f_{W|N}(w|n)$. (In words $f_{W|N}(w|n)$ is the pdf of W *given* that $N = n$.)

2.17 The time-to-failure in months, X, of light bulbs produced at two manufacturing plants A and B obey, respectively, the following CDFs

$$F_X(x) = (1 - e^{-x/5})u(x) \text{ for plant } A \tag{2.7-21}$$

$$F_X(x) = (1 - e^{-x/2})u(x) \text{ for plant } B. \tag{2.7-22}$$

Plant B produces two times as many bulbs as plant A. The bulbs, indistinguishable to the eye, are intermingled and sold. What is the probability that a bulb purchased at random will burn at least (a) two months; (b) five months; (c) seven months?

2.18 A smooth-surface table is ruled with equidistant parallel lines, a distance D apart. A needle of length L, where $L \leq D$, is dropped onto the table. What is the probability that the needle will intersect one of the lines?

2.19 It has been found that the number of people Y waiting in a queue in the bank on payday obeys the Poisson law as

$$P[Y = k|X = x] = e^{-x}\frac{x^k}{k!}, \quad k \geq 0, x > 0$$

given that the normalized serving time of the teller x (i.e., the time it takes the teller to deal with a customer) is constant. However, the serving time is more accurately modeled as an RV X. For simplicity let X be a uniform RV with

$$f_X(x) = \frac{1}{5}[u(x) - u(x-5)].$$

Then $P[Y = k|X = x]$ is still Poisson but $P[Y = k]$ is something else. Compute $P[Y = k]$ for $k = 0, 1,$ and 2. The answer for general k may be difficult.

2.20 Suppose in a presidential election each vote has equal probability $p = 0.5$ of being in favor of either of two candidates, candidate 1 and candidate 2. Assume all votes are independent. Suppose 8 votes are selected for inspection. Let X be the random variable that represents the number of favorable votes for candidate 1 in these 8 votes. Let A be the event that this number of favorable votes exceeds 4, that is, $A = \{X > 4\}$.

(a) What is the PMF for the random variable X ? Note that the PMF should be symmetric about $X = k = 4$.
(b) Find and plot the conditional distribution function $F_X(x|A)$ for the range $-1 \leq x \leq 10$.
(c) Find and plot the conditional pdf $f_X(x|A)$ for the range $-1 \leq x \leq 10$.
(d) Find the conditional probability that the number of favorable votes for candidate 1 is between 4 and 5 inclusive, that is, $P[4 \leq X \leq 5|A]$.

2.21 Suppose that random variables X and Y have a joint density function

$$f(x, y) = \begin{cases} A(2x + y), & 2 < x < 6, 0 < y < 5 \\ 0, & \text{otherwise} \end{cases}$$

Find

(a) the constant A
(b) $P(X>3)$
(c) $F_y(2)$
(d) $P(3 < X < 4/Y > 2)$

2.22 Consider the joint pdf of X and Y:

$$f_{XY}(x, y) = \frac{1}{3\pi} e^{-\frac{1}{2}[(x/2)^2 + (y/3)^2]} u(x)u(y).$$

Are X and Y independent RVs? Compute the probability of $\{0 < X \leq 2, 0 < Y \leq 3\}$.

2.23 The radial miss distance (in meters/m^2) of the landing point of a parachuting sky diver from the centre of the target area is known to be a Rayleigh random variable X with parameter $\sigma^2 = 100$.

(a) Find the radius r such that $P\ (X > r) = e^{-1}$
(b) Find the probability of the sky diver landing within a 10m radius from the centre of the target area, given that the landing is within 50m from the centre of the target area.

2.24 Show that Equation 2.6-75 factors as $f_X(x)f_Y(y)$ when $\rho = 0$. What are $f_X(x)$ and $f_Y(y)$? For $\sigma = 1$ and $\rho = 0$, what is $P[-\frac{1}{2} < X \le \frac{1}{2}, -\frac{1}{2} < Y \le \frac{1}{2}]$?

***2.25** Consider a communication channel corrupted by noise. Let X be the value of the transmitted signal and Y be the value of the received signal. Assume that the conditional density of Y given $\{X = x\}$ is Gaussian, that is,

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-x)^2}{2\sigma^2}\right),$$

and X is uniformly distributed on $[-1, 1]$. What is the conditional pdf of X given Y, that is, $f_{X|Y}(x|y)$?

2.26 Consider a communication channel corrupted by noise. Let X be the value of the transmitted signal and Y the value of the received signal. Assume that the conditional density of Y given X is Gaussian, that is,

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right),$$

and that X takes on only the values $+1$ and -1 equally likely. What is the conditional density of X given Y, that is, $f_{X|Y}(x|y)$?

2.27 The arrival time of a professor to his office is a continuous RV uniformly distributed over the hour between 8 A.M. and 9 A.M. Define the events:

$$A = \{\text{The prof. has not arrived by 8.30 A.M.}\}, \quad (2.7\text{-}23)$$

$$B = \{\text{The prof. will arrive by 8:31 A.M.}\}. \quad (2.7\text{-}24)$$

Find

(a) $P[B|A]$.
(b) $P[A|B]$.

2.28 Let X be a random variable with *pdf*

$$f_X(x) = \begin{cases} 0, & x < 0, \\ ce^{-2x}, & x \ge 0, \end{cases} \qquad (c > 0).$$

(a) Find c;
(b) Let $a > 0$, $x > 0$, find $P[X \ge x + a]$;
(c) Let $a > 0$, $x > 0$, find $P[X \ge x + a | X \ge a]$.

2.29 To celebrate getting a passing grade in a course on probability, Wynette invites her Professor, Dr. Chance, to dinner at the famous French restaurant C'est Tres Chere. The probability of getting a reservation if you call y days in advance is given by $1 - e^{-y}$, where $y \ge 0$. What is the minimum numbers of days that Wynette should call in advance in order to have a probability of at least 0.90 of getting a reservation?

2.30 A U.S. defense radar scans the skies for unidentified flying objects (UFOs). Let M be the event that a UFO is present and M^c the event that a UFO is absent. Let $f_{X/M}(x|M) = \frac{1}{\sqrt{2\pi}} \exp(-0.5[x-r]^2)$ be the conditional pdf of the radar return signal X when a UFO is actually there, and let $f_{X/M}(x/M^c) = \frac{1}{\sqrt{2\pi}} \exp(-0.5[x]^2)$ be the conditional pdf of the radar return signal X when there is no UFO. To be specific, let $r = 1$ and let the *alert level* be $x_A = 0.5$. Let A denote the event of an alert, that is, $\{X > x_A\}$. Compute $P[A|M]$, $P[A^c|M]$, $P[A|M^c]$, $P[A^c|M^c]$.

2.31 In the previous problem assume that $P[M] = 10^{-4}$. Compute

$$P[M|A], P[M|A^c], P[M^c|A], P[M^c|A^c]. \text{ Repeat for } P[M] = 10^{-6}.$$

Note: By assigning drastically different numbers to $P[M]$, this problem attempts to illustrate the difficulty of using probability in some types of problems. Because a UFO appearance is so rare (except in Roswell, New Mexico), it may be considered a *one-time event* for which accurate knowledge of the prior probability $P[M]$ is near impossible. Thus, in the surprise attack by the Japanese on Pearl Harbor in 1941, while the radar clearly indicated a massive cloud of incoming objects, the signals were ignored by the commanding officer (CO). Possibly the CO assumed that the prior probability of an attack was so small that a radar failure was more likely.

***2.32** (*research problem: receiver-operating characteristics*) In Problem, $P[A|M^c]$ is known as α, the *probability of a false alarm*, while $P[M|A]$ is known as β, the *probability of a correct detection*. Clearly $\alpha = \alpha(x_A)$, $\beta = \beta(x_A)$. Write a MATLAB program to plot β *versus* α for a fixed value of r. Choose $r = 0, 1, 2, 3$. The curves so obtained are known among radar people as the *receiver-operating characteristic* (ROC) for various values of r.

2.33 A sophisticated house security system uses an infrared beam to complete a circuit. If the circuit is broken, say by a robber crossing the beam, a bell goes off. The way the system works is as follows: The photodiode generates a beam of infrared photons at a Poisson rate of 8×10^6 photons per second. Every microsecond a counter counts the total number of photons collected at the detector. If the count drops below 2 photons in the counting interval (10^{-6} seconds), it is assumed that the circuit is broken and the bell rings. Assuming the Poisson PMF, compute the probability of a false alarm during a one-second interval.

2.34 A traffic light can be in one of three states: *green* (G), *red* (R), and *yellow* (Y). The light changes in a random fashion (e.g., the light at the corner of Hoosick and Doremus in Nirvana, New York). At any one time the light can be in only one state. The experiment consists of observing the state of the light.

(a) Give the sample space of this experiment and list five events.

(b) Let a random variable $X(\cdot)$ be defined as follows: $X(\text{G}) = -1$; $X(\text{R}) = 0$; $X(\text{Y}) = \pi$. Assume that $P[\text{G}] = P[\text{Y}] = 0.5 \times P[\text{R}]$. Plot the pdf of X. What is $P[X \leq 3]$?

***2.35** A *token-based, multi-user* communication system works as follows: say that nine user-stations are connected to a ring and an electronic signal, called a *token*, is passed around the ring in, say, a clockwise direction. The token stops at each station and

allows the user (if there is one) up to five minutes of signaling a message. The token waits for a maximum of one minute at each station for a user to initiate a message. If no user appears at the station at the end of the minute, the token is passed on to the next station. The five-minute window *includes* the waiting time of the token at the station. Thus, a user who begins signaling at the end of the token waiting period has only four minutes of signaling left.

(a) Assume that you are a user at a station. What are the minimum and maximum waiting times you might experience? The token is assumed to travel instantaneously from station to station.

(b) Let the probability that a station is occupied be p. If a station is occupied, the "occupation time" is a random variable that is uniformly distributed in $(0, 5)$ minutes. Using MATLAB, write a program that simulates the waiting time at your station. Assume that the token has just left your station. Pick various values of p.

2.36 All manufactured devices and machines fail to work sooner or later. Suppose that the failure rate is constant and the time to failure (in hours) is an exponential random variable X with parameter λ.

(a) Measurements show that the probability that the time to failure for computer memory chips in a given class exceeds 10^{-4} is e^{-1}. Calculate λ.

(b) Using the above value of λ, calculate the time x_0 such that 0.05 is the probability that the time to failure is less than x_0.

2.37 We are given the following joint pdf for random variables X and Y:

$$f_{X,Y}(x,y) = \begin{cases} A, & 0 \leq |x| + |y| < 1, \\ 0, & \text{else.} \end{cases}$$

(a) What is the value of the constant A?

(b) What is the marginal density $f_X(x)$?

(c) Are X and Y independent? Why?

(d) What is the conditional density $f_{Y|X}(y|x)$?

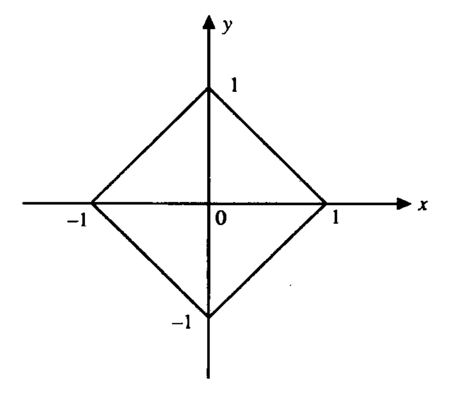


Figure P2.37 Support of $f_{XY}(x, y)$ in problem 2.37.

2.38 A laser used to scan the bar code on supermarket items is assumed to have a constant conditional failure rate $\lambda\,(>0)$. What is the maximum value of λ that will yield a probability of a first breakdown in 100 hours of operation less than or equal to 0.05?

2.39 Compute the pdf of the failure time X if the conditional failure rate $\alpha(t)$ is as shown in Figure P2.39.

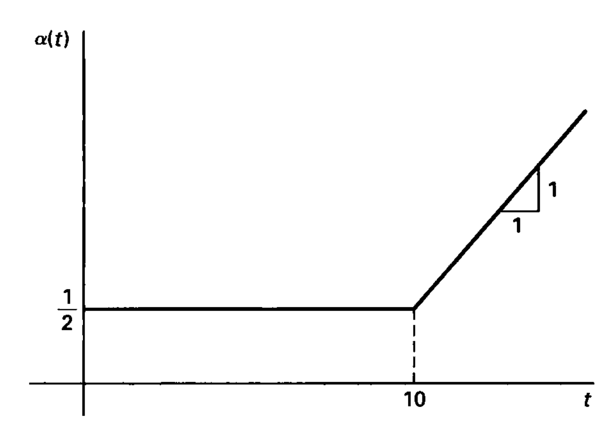


Figure P2.39 Failure rate $\alpha(t)$ in problem 2.39.

2.40 Two people agree to meet between 2.00 p.m. and 3.00 p.m, with the understanding that each will wait no longer than 15 minutes for the other. What is the probability that they will meet?

REFERENCES

2-1. M. Loeve, *Probability Theory.* New York: Van Nostrand, Reinhold, 1962.
2-2. W. Feller, *An Introduction to Probability Theory and Its Applications*, 2 vols. New York: John Wiley, 1950, 1966.
2-3. W. F. Davenport, *Probability and Random Processes: An Introduction for Applied Scientists and Engineers.* New York: McGraw-Hill, 1970.
2-4. B. Saleh, *Photoelectron Statistics.* New York: Springer-Verlag, 1978.
2-5. M. Born and E. Wolf, *Principles of Optics.* New York: Pergamon Press, 1965.
2-6. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions.* New York: Dover, 1965. (On-line at various locations.)

ADDITIONAL READING

Cooper, G. R. and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, 3rd edition. New York: Holt, Rinehart and Winston, 1999.

Leon-Garcia, A., *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd edition. Reading, MA: Prentice Hall, 2008.

Helstrom, C. W., *Probability and Stochastic Processes for Engineers*, 2nd edition. New York: Macmillan, 1991.

A. Papoulis, and S. U. Pillai *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 4th Ed, 2002.

Pebbles, P. Z. Jr., *Probability, Random Variables, and Random Signal Principles*, 4th edition. New York: McGraw-Hill, 2001.

Scheaffer, R. L., *Introduction to Probability and Its Applications.* Belmont, CA: Duxbury 1990.

Yates, R. D. and D. J. Goodman, *Probability and Stochastic Processes*, 2nd edition, New York: Wiley, 2004.

Ziemer, R. E., *Elements of Engineering Probability & Statistics.* Upper Saddle River, NJ: Prentice Hall, 1997.

3 Functions of Random Variables

3.1 INTRODUCTION

A classic problem in engineering is the following: We are given the input to a system and we must calculate the output. If the input to a system is random, the output will generally be random as well. To put this somewhat more formally, if the input at some instant t or point x is a random variable (RV), the output at some corresponding instant t' or point x' will be a random variable. Now the question arises, if we know the CDF, PMF, or pdf of the input RV can we compute these functions for the output RV? In many cases we can, while in other cases the computation is too difficult and we settle for descriptors of the output RV which contain less information than the CDF. Such descriptors are called *averages* or *expectations* and are discussed in Chapter 4. In general for systems with memory, that is, systems in which the output at a particular instant of time depends on past values of the input (possibly an infinite number of such past values), it is much more difficult (if not impossible) to calculate the CDF of the output. This is the case for random sequences and processes to be treated in Chapters 7 and 8. In this chapter, we study much simpler situations involving just one or a few random variables. We illustrate with some examples.

Example 3.1-1

(*power loss in resistor*) As is well known from electric circuit theory, the current I flowing through a resistor R (Figure 3.1-1) dissipates an amount of power W given by

$$W \triangleq W(I) = I^2 R. \tag{3.1-1}$$

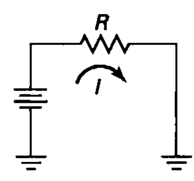


Figure 3.1-1 Ohmic power dissipation in a resistor.

Equation 3.1-1 is an explicit rule that generates for every value of I a number $W(I)$. This rule or correspondence is called a *function* and is denoted by $W(\cdot)$ or merely W or sometimes even $W(I)$—although the latter notation obscures the difference between the rule and the actual number. Clearly, if I were a random variable, the rule $W = I^2R$ generates a new random variable W whose CDF might be quite different from that of I.† Indeed, this alludes to the heart of the problem: Given a rule $g(\cdot)$, and a random variable X with pdf $f_X(x)$, what is the pdf $f_Y(y)$ of the random variable $Y = g(X)$?

The computation of $f_Y(y)$, $F_Y(y)$, or the PMF of Y, that is, $P_Y(y_i)$, can be very simple or quite complex. We illustrate such a computation with a second example, one that comes from communication theory.

Example 3.1-2

(*waveform detector*) A two-level waveform is made analog because of the effect of additive Gaussian noise (Figure 3.1-2). A decoder samples the analog waveform $x(t)$ at t_0 and decodes according to the following rule:

Input to Decoder x	Output of Decoder y
If $x(t_0)$:	Then y is assigned:
$\geq\frac{1}{2}$	1
$<\frac{1}{2}$	0

What is the PMF or pdf of Y?

Solution Clearly with Y (an RV) denoting the output of the decoder, we can write the following events:

$$\{Y = 0\} = \{X < 0.5\} \tag{3.1-2a}$$

$$\{Y = 1\} = \{X \geq 0.5\}, \tag{3.1-2b}$$

where $X \triangleq x(t_0)$. Hence if we assume $X\colon N(1, 1)$, we obtain the following:

$$P[Y = 0] = P[X < 0.5] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0.5} e^{-1/2(x-1)^2 dx}$$

$$\simeq 0.31. \tag{3.1-3}$$

†This is assuming that the composite function $I^2(\zeta)R$ satisfies the required properties of an RV (see Section 2.2).

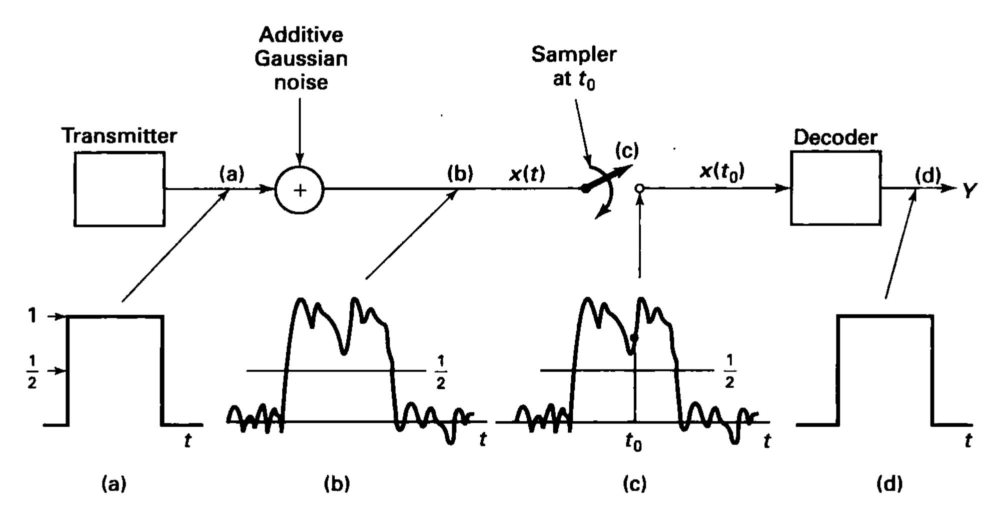


Figure 3.1-2 Decoding of a noise-corrupted digital pulse by sampling and hard clipping.

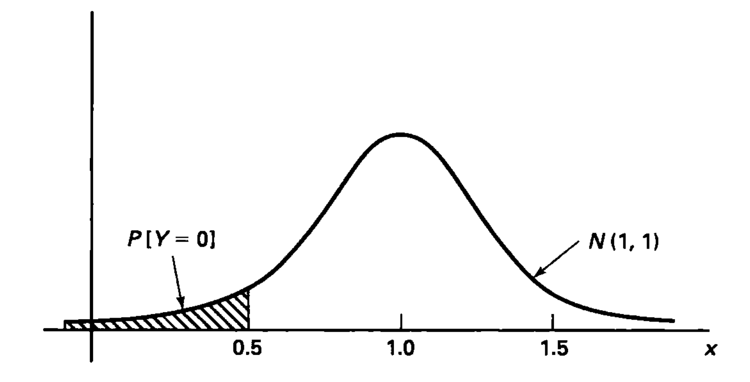


Figure 3.1-3 The area associated with $P[Y = 0]$ in Example 3.1-2.

In arriving at Equation 3.1-3 we use the normalization procedure explained in Section 2.4 and the fact that for $X\colon N(0,1)$ and any $x < 0$, the CDF $F_X(x) = \frac{1}{2} - \text{erf}(|x|)$. The area under the Normal $N(1,1)$ curve associated with $P[Y = 0]$ is shown in Figure 3.1-3.

In a similar fashion we compute $P[Y = 1] = 0.69$. Hence the PMF of Y is

$$P_Y(y) = \begin{cases} 0.31, \; y = 0, \\ 0.69, \; y = 1, \\ 0, \quad \text{else}. \end{cases} \tag{3.1-4}$$

Using Dirac delta functions, we can obtain the pdf of Y:

$$f_Y(y) = 0.31\,\delta(y) + 0.69\,\delta(y-1). \tag{3.1-5}$$

In terms of the Kronecker delta function, that is, $\delta(y) = 1$ at $y = 0$, equal 0 else, the PMF would be

$$P_Y(y) = 0.31\,\delta(y) + 0.69\,\delta(y-1).$$

The Knonecker δ is used in PMFs while the Dirac δ is used in pdffs. We keep the symbols the same although they mean different things.

Of course not all *function-of-a-random-variable* (FRV) problems are this easy to evaluate. To gain a deeper insight into the FRV problem, we take a closer look at the underlying concept of FRV. The gain in insight will be useful when we discuss random sequences and processes beginning in Chapter 7.

Functions of a Random Variable (FRV): Several Views

There are several different but essentially equivalent views of an FRV. We will now present two of them. The differences between them are mainly ones of *emphasis*.

Assume as always an underlying probability space $\mathscr{P} = (\Omega, \mathscr{F}, P)$ and a random variable X defined on it. Recall that X is a rule that assigns to every $\zeta \in \Omega$ a number $X(\zeta)$. X transforms the σ-field of events $\mathscr{F}$ into the Borel σ-field $\mathscr{B}$ of sets of numbers on the real line. If R_X denotes the subset of the real line actually reached by X as ζ roams over Ω, then we can regard X as an ordinary function with domain Ω and range R_X. Now, additionally, consider a measurable real function $g(x)$ of the real variable x.

First view ($Y\colon \Omega \longrightarrow R_Y$). For every $\zeta \in \Omega$, we generate a number $g(X(\zeta)) \triangleq Y(\zeta)$. The rule Y, which generates the numbers $\{Y(\zeta)\}$ for random outcomes $\{\zeta \in \Omega\}$, is an RV with domain Ω and range $R_Y \subset R^1$. Finally for every Borel set of real numbers B_Y, the set $\{\zeta\colon Y(\zeta) \in B_Y\}$ is an event. In particular the event $\{\zeta\colon Y(\zeta) \leq y\}$ is equal to the event $\{\zeta\colon g(X(\zeta)) \leq y\}$.

In this view, the stress is on Y as a mapping from Ω to R_Y. The intermediate role of X is suppressed.

Second view (input/output systems view). For every value of $X(\zeta)$ in the range R_X, we generate a new number $Y = g(X)$ whose range is R_Y. The rule Y whose domain is R_X and range is R_Y is a function of the random variable X. In this view the stress is on viewing Y as a mapping from one set of real numbers to another. A model for this view is to regard X as the input to a system with transformation function $g(\cdot)$.[†] For such a system, an input x gets transformed to an output $y = g(x)$ and an input function X gets transformed to an output function $Y = g(X)$. (See Figure 3.1-4.)

The input–output viewpoint is the one we stress, partly because it is particularly useful in dealing with random processes where the input consists of waveforms or sequences of random variables. The central problem in computations involving FRVs is: Given $g(x)$ and $F_X(x)$, find the point set C_y such that the following events are equal:

[†] g can be any *measurable* function; that is, if R_Y is the range of Y, then the inverse image (see Section 2.2) of every subset in R_Y generated by countable unions and intersections of sets of the form $\{Y \leq y\}$ is an event.

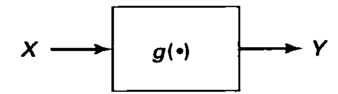


Figure 3.1-4 Input/output view of a function of a random variable.

$$\{\zeta : Y(\zeta) \leq y\} = \{\zeta : g[X(\zeta)] \leq y\}$$
$$= \{\zeta : X(\zeta) \in C_y\}. \quad (3.1\text{-}6a)$$

In general we will economize on notation and write Eq. 3.1-6a as $\{Y \leq y\} = \{X \in C_y\}$ in the sequel. For C_y so determined it follows that

$$P[Y \leq y] = P[X \in C_y] \quad (3.1\text{-}6b)$$

since the underlying event is the same. If C_y is empty, then the probability of $\{Y \leq y\}$ is zero.

In dealing with the input–output model, it is generally convenient to omit any references to an abstract underlying experiment and deal, instead, directly with the RVs X and Y. In this approach the underlying experiments are the observations on X, events are Borel subsets of the real line R^1, and the set function $P[\cdot]$ is replaced by the distribution function $F_X(\cdot)$. Then Y is a mapping (an RV) whose domain is the range R_X of X, and whose range R_Y is a subset of R^1. The functional properties of X are ignored in favor of viewing X as a mechanism that gives rise to numerically valued random phenomena. In this view the domain of X is irrelevant.

Additional discussion on the various views of an FRV are available in the literature.†

3.2 SOLVING PROBLEMS OF THE TYPE $Y = g(X)$

We shall now demonstrate how to solve problems of the type $Y = g(X)$. Eventually we shall develop a formula that will enable us to solve problems of this type very rapidly. However, use of the formula at too early a stage of the development will tend to mask the underlying principles needed to deal with more difficult problems.

Example 3.2-1

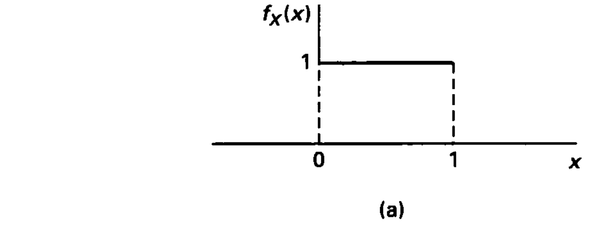
Let X be a uniform RV on $(0, 1)$, that is, $X{:}\,U(0, 1)$, and let $Y = 2X + 3$. Then we need to find the point set C_y in Equation 3.1-6b to compute $F_Y(y)$. Clearly

$$\{Y \leq y\} = \{2X + 3 \leq y\} = \left\{X \leq \tfrac{1}{2}(y - 3)\right\}.$$

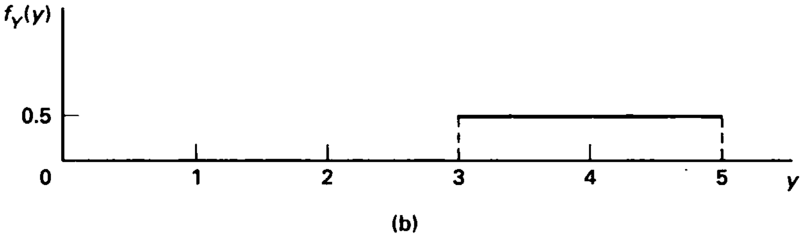
Hence C_y is the interval $(-\infty, \frac{1}{2}(y - 3))$ and

$$F_Y(y) = F_X\left(\frac{y-3}{2}\right).$$

†For example see Davenport [3-1, p.174] or Papoulis and Pillai [3-5].



(a)



(b)

Figure 3.2-1 (a) Original pdf of X; (b) the pdf of $Y = 2X + 3$.

The pdf of Y is

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\left[F_X\left(\frac{y-3}{2}\right)\right] = \frac{1}{2}f_X\left(\frac{y-3}{2}\right).$$

The solution is shown in Figure 3.2-1.

Generalization. Let $Y = aX + b$ with X a continuous RV with pdf $f_X(x)$. Then for $a > 0$ the outcomes $\{\zeta\} \subset \Omega$ that produce the event $\{aX + b \leq y\}$ are identical with the outcomes $\{\zeta\} \subset \Omega$ that produce the event $\{X \leq \frac{y-b}{a}\}$. Thus,

$$\{Y \leq y\} = \{aX + b \leq y\} = \left\{X \leq \frac{y-b}{a}\right\}.$$

From the definition of the CDF:

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right), \tag{3.2-1}$$

and so

$$f_Y(y) = \frac{1}{a}f_X\left(\frac{y-b}{a}\right). \tag{3.2-2}$$

For $a < 0$, the following events are equal†

$$\{Y \leq y\} = \{aX + b \leq y\} = \left\{X \geq \frac{y-b}{a}\right\}.$$

†By which we mean that the event $\{\zeta\colon Y(\zeta) \leq y\} = \left\{\zeta\colon X(\zeta) \geq \frac{y-b}{a}\right\}$.

Since the events $\left\{X < \frac{y-b}{a}\right\}$ and $\left\{X \geq \frac{y-b}{a}\right\}$ are disjoint and their union is the certain event, we obtain from Axiom 3

$$P\left[X < \frac{y-b}{a}\right] + P\left[X \geq \frac{y-b}{a}\right] = 1.$$

Finally for a continuous RV

$$P\left[X < \frac{y-b}{a}\right] = P\left[X \leq \frac{y-b}{a}\right]$$

and

$$P\left[X \geq \frac{y-b}{a}\right] = P\left[X > \frac{y-b}{a}\right].$$

Thus, for $a < 0$

$$F_Y(y) = 1 - F_X\left(\frac{y-b}{a}\right) \tag{3.2-3}$$

and

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right), \qquad a \neq 0. \tag{3.2-4}$$

When X is not necessarily continuous, we would have to modify the development in the case $a < 0$ because it may no longer be true that $P\left[X < \frac{y-b}{a}\right] = P\left[X \leq \frac{y-b}{a}\right]$ because of the possibility that the event $\{X = \frac{y-b}{a}\}$ has a positive probability. The modified statement then becomes $P\left[X < \frac{y-b}{a}\right] = P\left[X \leq \frac{y-b}{a}\right] - P[X = \frac{y-b}{a}] = F_X\left(\frac{y-b}{a}\right) - P_X(\frac{y-b}{a})$, where we have employed the PMF P_X to subtract the probability of this event. The final answer for the case $a < 0$ must be changed accordingly.

Example 3.2-2

(*square-law detector*) Let X be an RV with continuous CDF $F_X(x)$ and let $Y \triangleq X^2$. Then

$$\{Y \leq y\} = \{X^2 \leq y\} = \{-\sqrt{y} \leq X \leq \sqrt{y}\} = \{-\sqrt{y} < X \leq \sqrt{y}\} \cup \{X = -\sqrt{y}\}. \tag{3.2-5}$$

The probability of the union of disjoint events is the sum of their probabilities. Using the definition of the CDF, we obtain

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) + P[X = -\sqrt{y}]. \tag{3.2-6}$$

If X is continuous, $P[X = -\sqrt{y}] = 0$. Then for $y > 0$,

$$f_Y(y) = \frac{d}{dy}[F_Y(y)] = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}). \tag{3.2-7}$$

For $y < 0$, $f_Y(y) = 0$. How do we know this? Recall from Equation 3.1-6a that if C_y is empty, then $P[Y \in C_y] = 0$ and hence $f_Y(y) = 0$. For $y < 0$, there are no values of the RV X on the real line that satisfy

$$\{-\sqrt{y} \leq X \leq \sqrt{y}\}.$$

Hence $f_Y(y) = 0$ for $y < 0$. If $X\colon N(0,1)$, then from Equation 3.2-7,

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y} u(y), \tag{3.2-8}$$

where $u(y)$ is the standard unit step function. Equation 3.2-7 is the Chi-square pdf with one degree-of-freedom.

Example 3.2-3

(*half-wave rectifier*) A half-wave rectifier has the transfer characteristic $g(x) = xu(x)$ (Figure 3.2-2).

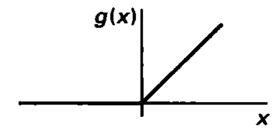


Figure 3.2-2 Half-wave rectifier.

Thus,

$$F_Y(y) = P[X u(X) \le y] = \int_{\{x\colon xu(x) \le y\}} f_X(x)dx. \tag{3.2-9}$$

(i) Let $y > 0$; then $\{x\colon xu(x) \le y\} = \{x\colon x > 0; x \le y\} \cup \{x\colon x \le 0\} = \{x\colon x \le y\}$. Thus $F_Y(y) = \int_{-\infty}^{y} f_X(x)dx = F_X(y)$.

(ii) Next let $y = 0$. Then $P[Y = 0] = P[X \le 0] = F_X(0)$.

(iii) Finally let $y < 0$. Then $\{x\colon xu(x) \le y\} = \phi$ (the empty set).

Thus,

$$F_Y(y) = \int_{\phi} f_X(x)dx = 0.$$

If $X\colon N(0,1)$, then $F_Y(y)$ has the form in Figure 3.2-3.

The pdf is obtained by differentiation. Because of the discontinuity of $y = 0$, we obtain a Dirac impulse in the pdf at $y = 0$, that is,

$$f_Y(y) = \begin{cases} 0, & y < 0, \\ F_X(0)\delta(y), & y = 0, \\ f_X(y), & y > 0. \end{cases} \tag{3.2-10}$$

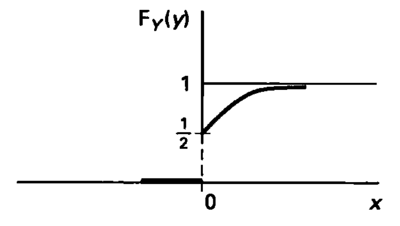


Figure 3.2-3 The CDF of Y when X: $N(0,1)$ for the half-wave rectifier.

This can be compactly written as $f_Y(y) = f_X(y)u(y) + F_X(0)\delta(y)$. We note that in this problem it is not true that $P[Y < 0] = P[Y \leq 0]$. There is a non-zero probability that $P[Y = 0]$.

Example 3.2-4

Let X be a Bernoulli RV with $P[X = 0] = p$ and $P[X = 1] = q$. Then

$$f_X(x) = p\delta(x) + q\delta(x-1) \text{ and } F_X(x) = pu(x) + qu(x-1),$$

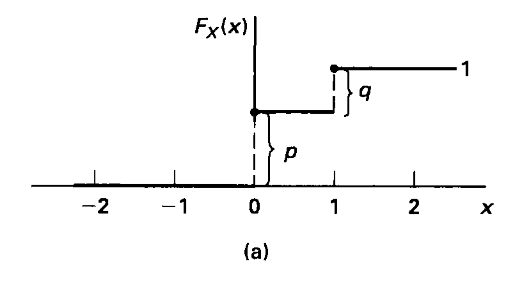
where $u(x)$ is the unit-step function of continuous variable x.

Let $Y \triangleq X - 1$. Then (Figure 3.2-4)

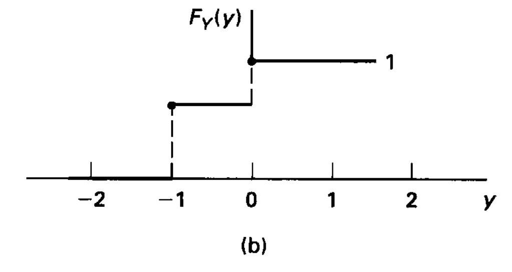
$$\begin{aligned} F_Y(y) &= P[X - 1 \leq y] \\ &= P[X \leq y + 1] \\ &= F_X(y+1) \\ &= pu(y+1) + qu(y). \end{aligned}$$

The pdf is

$$f_Y(y) = \frac{d}{dy}[F_Y(y)] = p\delta(y+1) + q\delta(y). \tag{3.2-11}$$



(a)



(b)

Figure 3.2-4 (a) CDF of X; (b) CDF of $Y = X - 1$.

Example 3.2-5

(*transformation of* CDFs) Let X have a continuous CDF $F_X(x)$ that is a *strict monotone* increasing function† of x. Let Y be an RV formed from X by the transformation with the CDF function itself,

$$Y = F_X(X). \tag{3.2-12}$$

To compute $F_Y(y)$, we proceed as usual:

$$\begin{aligned}\{Y \leq y\} &= \{F_X(X) \leq y\} \\ &= \{X \leq F_X^{-1}(y)\}.\end{aligned}$$

Hence

$$\begin{aligned}F_Y(y) &= P[F_X(X) \leq y] \\ &= P[X \leq F_X^{-1}(y)] \\ &= \int_{\{x:\ F_X(x) \leq y\}} f_X(x)dx.\end{aligned}$$

1. Let $y < 0$. Then since $0 \leq F_X(x) \leq 1$ for all $x \in [-\infty, \infty]$, the set $\{x\colon F_X(x) \leq y\} = \phi$ and $F_Y(y) = 0$.
2. Let $y > 1$. Then $\{x\colon F_X(x) \leq y\} = [-\infty, \infty]$ and $F_Y(y) = 1$.
3. Let $0 \leq y \leq 1$. Then $\{x\colon F_X(x) \leq y\} = \{x\colon x \leq F_X^{-1}(y)\}$

and

$$F_Y(y) = \int_{-\infty}^{F_X^{-1}(y)} f_X(x)dx = F_X(F_X^{-1}(y)) = y.$$

Hence

$$F_Y(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \leq y \leq 1, \\ 1, & y > 1. \end{cases} \tag{3.2-13}$$

Equation 3.2-13 says that whatever probability law X obeys, so long as it is continuous and strictly monotonic, $Y \triangleq F_X(X)$ will be a *uniform.* Conversely, given a uniform distribution for Y, the transformation $X \triangleq F_X^{-1}(Y)$ will generate an RV with contiuous and strictly monotonic CDF $F_X(x)$ (Figure 3.2-5). This technique is sometimes used in *simulation* to generate RVs with specified distributions from a uniform RV.

Example 3.2-6

(*transform uniform to standard Normal*) From the last example, we can transform a uniform RV $X{:}U[0,1]$ to any continuous distribution that has a strictly increasing CDF. If we want a standard Normal, that is, Gaussian $Y{:}N(0,1)$, its CDF is given as

†In other words $x_2 > x_1$ implies $F_X(x_2) > F_X(x_1)$, that is, without the possibility of equality.

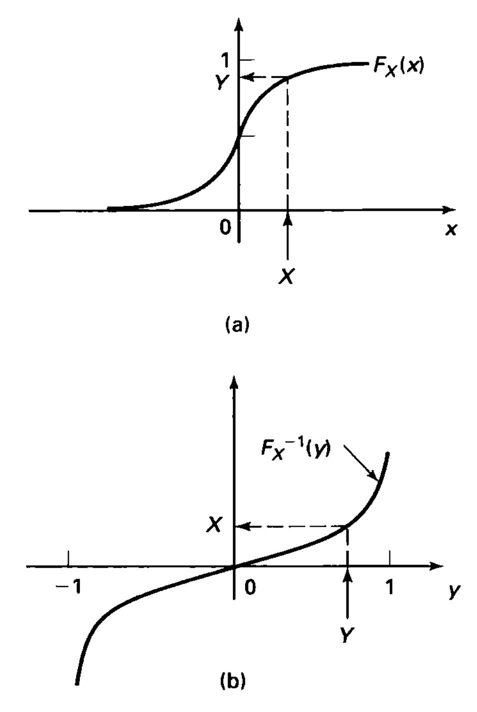


Figure 3.2-5 Generating an RV with CDF $F_X(x)$ from a uniform RV. (a) Creating a uniform RV Y from an RV X with CDF $F_X(x)$; (b) creating an RV with CDF $F_X(x)$ from a uniform RV Y.

$$F_Y(y) = \begin{cases} \frac{1}{2} + \operatorname{erf}(y),\ y \geq 0, \\ \frac{1}{2} - \operatorname{erf}(-y),\ y < 0. \end{cases}$$

Solving for the required transformation $g(x)$, we get

$$g(x) = \begin{cases} -\operatorname{erf}^{-1}(\frac{1}{2} - x),\ 0 \leq x < \frac{1}{2}, \\ \operatorname{erf}^{-1}(x - \frac{1}{2}),\ \frac{1}{2} \leq x \leq 1. \end{cases}$$

A plot of this transformation is given in Figure 3.2-6.

A MATLAB program `transformCDF.m` available at the book website can be used to generate relative frequency histograms of this transformation in action. The following results were obtained with 1000 trials. Figure 3.2-7 shows the histogram of the 1000 RVs distributed as $U[0, 1]$. Figure 3.2-8b shows the corresponding histogram of the transformed RVs.

Example 3.2-7

(*quantizing*) In analog-to-digital conversion, an analog waveform is sampled, quantized, and coded (Figure 3.2-9). A quantizer is a function that assigns to each sample x, a value from a set $Q \triangleq \{y_{-N}, \ldots, y_0, \ldots, y_N\}$ of $2N + 1$ predetermined values [3-2]. Thus, an uncountably

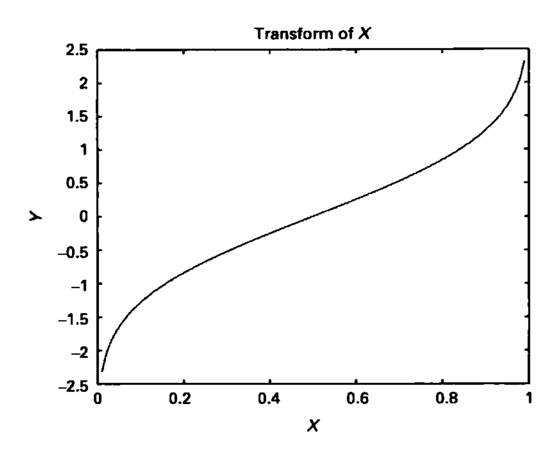


Figure 3.2-6 Plot of transformation $y = g(x) = F_{SN}^{-1}(x)$ that transforms $U[0,1]$ into $N(0,1)$.

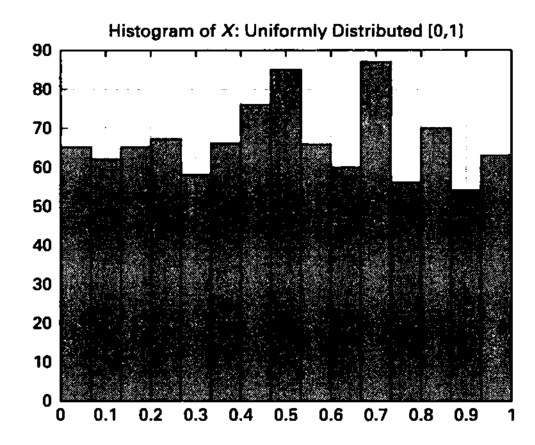


Figure 3.2-7 Histogram of 1000 i.i.d. RVs distributed as $U[0,1]$.

infinite set of values (the analog input x) is reduced to a finite set (some digital output y_i). Note that this practical quantizer is also a limiter, that is, for x greater than some y_N or less than some y_{-N}, the output is $y = y_N$ or y_{-N}, respectively.

A common quantizer is the uniform quantizer, which is a staircase function of uniform step size a, that is,

$$g(x) = ia \qquad (i-1)a < x \leq ia, \qquad i \text{ an integer.} \tag{3.2-14}$$

Thus, the quantizer assigns to each x the closest value of ia above continuous sample value $3x$ as is shown by the staircase function in Figure 3.2-10.

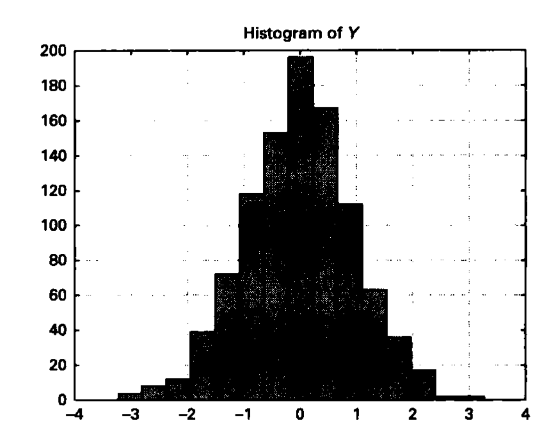


Figure 3.2-8 Histogram of 1000 transformed i.i.d. RVs.

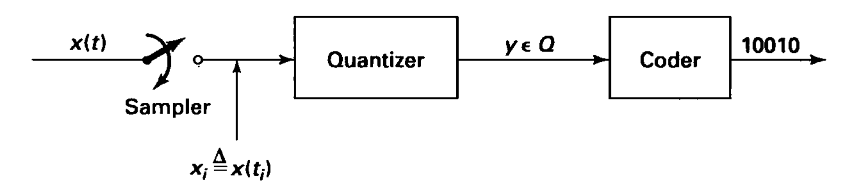


Figure 3.2-9 An analog-to-digital converter.

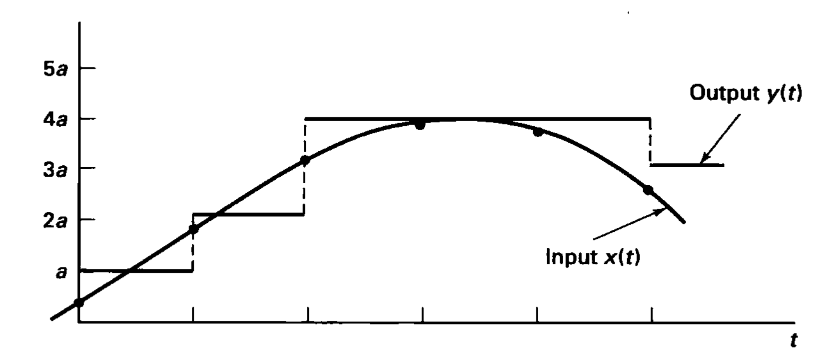


Figure 3.2-10 Quantizer output (staircase function) versus input (continuous line).

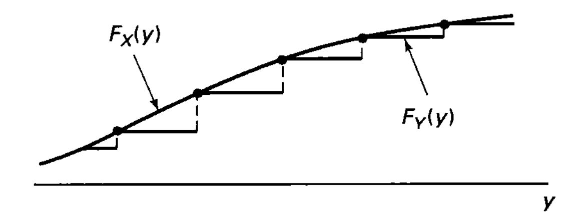


Figure 3.2-11 $F_X(y)$ versus $F_Y(y)$.

If X is an RV denoting the sampled value of the input and Y denotes the quantizer output, then with $a = 1$, we get the output PMF

$$\begin{aligned} P_Y(i) &= P[Y = i] \\ &= P[i - 1 < X \leq i] \\ &= F_X(i) - F_X(i - 1). \end{aligned}$$

The output CDF then becomes the staircase function

$$\begin{aligned} F_Y(y) &= \sum_i P_Y(i)u(y - i) \\ &= \sum_i [F_X(i) - F_X(i - 1)]u(y - i), \end{aligned} \tag{3.2-15}$$

as sketched in Fig. 3.2-12.

When $y = n$ (an integer), $F_Y(n) = F_X(n)$, otherwise $F_Y(y) < F_X(y)$.

Example 3.2-8

(*sine wave*) A classic problem is to determine the pdf of $Y = \sin X$, where $X : U(-\pi, +\pi)$, that is, uniformly distributed over $(-\pi, +\pi)$. From Figure 3.2-12 we see that for $0 \leq y \leq 1$, the event $\{Y \leq y\}$ satisfies

$$\begin{aligned} \{Y \leq y\} &= \{\sin X \leq y\} \\ &= \{-\pi < X \leq \sin^{-1} y\} \cup \{\pi - \sin^{-1} y < X \leq \pi\}. \end{aligned}$$

Since the two events on the last line are disjoint, we obtain

$$F_Y(y) = F_X(\pi) - F_X(\pi - \sin^{-1} y) + F_X(\sin^{-1} y) - F_X(-\pi). \tag{3.2-16}$$

Hence

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= f_X(\pi - \sin^{-1} y)\frac{1}{\sqrt{1 - y^2}} + f_X(\sin^{-1} y)\frac{1}{\sqrt{1 - y^2}} \end{aligned} \tag{3.2-17}$$

$$= \frac{1}{2\pi}\frac{1}{\sqrt{1-y^2}} + \frac{1}{2\pi}\frac{1}{\sqrt{1-y^2}} \tag{3.2-18}$$

$$= \frac{1}{\pi}\frac{1}{\sqrt{1-y^2}} \qquad 0 \leq y < 1. \tag{3.2-19}$$

If this calculation is repeated for $-1 < y \leq 0$, the diagram in Figure 3.2-12 changes to that of Figure 3.2-13. So the event $\{Y \leq y\} = \{\sin X \leq y\}$ expressed in terms of the RV X becomes

$$\{-\pi - \sin^{-1} y < X \leq \sin^{-1} y\}, \tag{3.2-20}$$

where we are now using the inverse sin appropriate for $y \leq 0$. Then we can write the following equation for the CDFs

$$F_Y(y) = F_X(\sin^{-1} y) - F_X(-\pi - \sin^{-1} y).$$

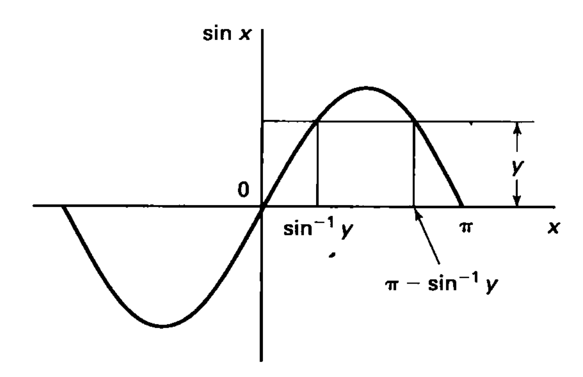


Figure 3.2-12 Graph showing roots of $y = \sin x$ when $0 \leq y < 1$.

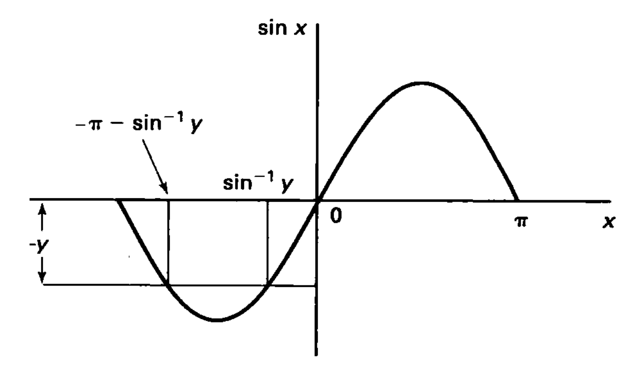


Figure 3.2-13 Plot showing roots of $y = \sin x$ when $-1 < y < 0$.

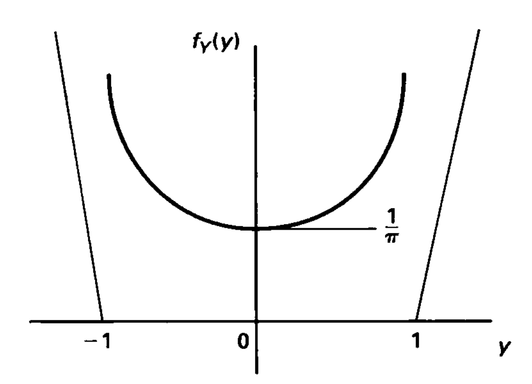


Figure 3.2-14 The probability density function of $Y = \sin X$.

Upon differentiation, we obtain

$$
\begin{aligned}
f_Y(y) &= \frac{dF_Y(y)}{dy} \\
&= f_X(\sin^{-1} y)\frac{1}{\sqrt{1-y^2}} - f_X(-\pi - \sin^{-1} y)\frac{1}{\sqrt{1-y^2}} \\
&= \frac{1}{2\pi}\frac{1}{\sqrt{1-y^2}} + \frac{1}{2\pi}\frac{1}{\sqrt{1-y^2}} \\
&= \frac{1}{\pi}\frac{1}{\sqrt{1-y^2}}, \qquad -1 < y \leq 0,
\end{aligned}
$$

which is the same form as before when $0 \leq y < 1$.

Finally we consider $|y| > 1$, and since $|\sin(x)| \leq 1$ for all x, we see that the pdf f_Y must be zero there. Combining these three results, we obtain the complete solution (Figure 3.2-14):

$$
f_Y(y) = \begin{cases} \frac{1}{\pi}\frac{1}{\sqrt{1-y^2}}, & |y| < 1, \\ 0, & \text{otherwise}. \end{cases} \tag{3.2-21}
$$

We shall now go on to derive a simple formula that will enable us to solve many problems of the type $Y = g(X)$ by going directly from pdf to pdf, without the need to find the CDF first. We shall call this new approach the *direct* method. For some problems, however, the *indirect* method of this past section may be less prone to error.

General Formula of Determining the pdf of $Y = g(X)$

We are given the continuous RV X with pdf $f_X(x)$ and the differentiable function $g(x)$ of the real variable x. What is the pdf of $Y \triangleq g(X)$?

Solution The event $\{y < Y \leq y + dy\}$ can be written as a union of disjoint elementary events $\{E_i\}$ in the Borel field generated under X. If the equation $y = g(x)$ has a finite number n of real roots† $x_1, \ldots, x_n$, then the disjoint events have the form $E_i = \{x_i - |dx_i| < X \leq x_i\}$ if $g'(x_i)$ is negative or $E_i = \{x_i < X \leq x_i + |dx_i|\}$ if $g'(x_i)$ is positive.‡ (See Figure 3.2-15.) In either case, it follows from the definition of the pdf that $P[E_i] = f_X(x_i)|dx_i|$. Hence

$$P[y < Y \leq y + dy] = f_Y(y)|dy|$$
$$= \sum_{i=1}^{n} f_X(x_i)|dx_i| \qquad (3.2\text{-}22)$$

or, equivalently, if we divide through by $|dy|$

$$f_Y(y) = \sum_{i=1}^{n} f_X(x_i) \left|\frac{dx_i}{dy}\right| = \sum_{i=1}^{n} f_X(x_i) \left|\frac{dy}{dx_i}\right|^{-1}.$$

At the roots of $y = g(x)$, $dy/dx_i = g'(x_i)$, and we obtain the important formula

$$f_Y(y) = \sum_{i=1}^{n} f_X(x_i)/|g'(x_i)| \qquad x_i = x_i(y), \qquad g'(x_i) \neq 0. \qquad (3.2\text{-}23)$$

Equation 3.2-23 is a fundamental equation that is very useful in solving problems where the transformation $g(x)$ has several roots. Note that we need to make the assumption that

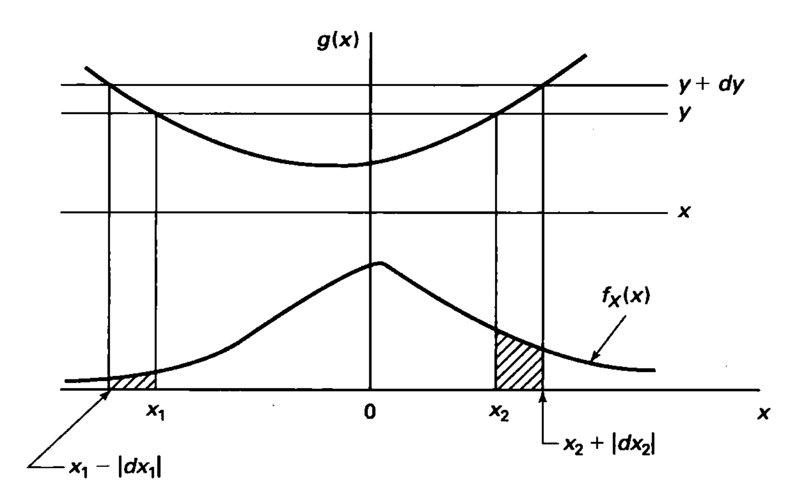


Figure 3.2-15 The event $\{y < Y \leq y + dy\}$ is the union of two disjoint events on the probability space of X.

†By roots we mean the set of points x_i such that $y - g(x_i) = 0$, $i = 1, \ldots, n$.
‡The prime indicates derivatives with respect to x.

$g'(x_i) \neq 0$ at all the roots. To see what happens otherwise, realize that a region where $g' = 0$ is a flat region for the transformation g. So for any x in this flat region, the y value is identical, and that will create a probability mass at this value of y whose amount is equal to the probability of the event that X falls in this flat region. In terms of the pdf f_Y, the mass would turn into an impulse with area equal to the mass.

If, for a given y, the equation $y - g(x) = 0$ has no real roots, then $f_Y = 0$ at that y.†
Figure 3.2-15 illustrates the case when $n = 2$.

Example 3.2-9

(*trig function of X*) To illustrate the use of Equation 3.2-23, we solve Example 3.2-8 by using this formula. Thus we seek the pdf of $Y = \sin X$ when the pdf of X is $f_X(x) = 1/2\pi$ for $-\pi < x \leq \pi$. Here the function g is $g(x) = \sin x$. The two roots of $y - g(x) = y - \sin x = 0$ for $y > 0$ are $x_1 = \sin^{-1} y$, $x_2 = \pi - \sin^{-1} y$. Also

$$\frac{dg}{dx} = \cos x,$$

which must be evaluated at the two roots x_1 and x_2. At $x_1 = \sin^{-1} y$ we get $dg/dx|_{x=x_1} = \cos(\sin^{-1} y)$. Likewise when $x_2 = \pi - \sin^{-1} y$, we get

$$\begin{aligned} \left.\frac{dg}{dx}\right|_{x=x_2} &= \cos(\pi - \sin^{-1} y) = \cos\pi \times \cos(\sin^{-1} y) + \sin\pi \times \sin(\sin^{-1} y) \\ &= -\cos(\sin^{-1} y). \end{aligned}$$

The quantity $\cos(\sin^{-1} y)$ can be further evaluated with the help of Figure 3.2-16. There we see that $\theta = \sin^{-1} y$ and $\cos\theta = \sqrt{1 - y^2} = \cos(\sin^{-1} y)$. Hence

$$\left|\frac{dg}{dx}\right|_{x_1} = \left|\frac{dg}{dx}\right|_{x_2} = \sqrt{1 - y^2}.$$

Finally, $f_X(\sin^{-1} y) = f_X(\pi - \sin^{-1} y) = 1/2\pi$. Using these results in Equation 3.2-23 enables us to write

$$f_Y(y) = \frac{1}{\pi}\frac{1}{\sqrt{1 - y^2}} \qquad 0 \leq y < 1,$$

which is the same result as in Equation 3.2-19. Repeating this procedure for $y < 0$ then gives the same solution for all y as is given in Equation 3.2-21.

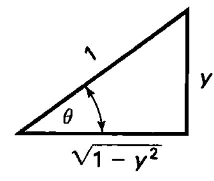


Figure 3.2-16 Evaluating $\cos(\sin^{-1} y)$.

†The RV X, being real valued, cannot take on values that are imaginary.

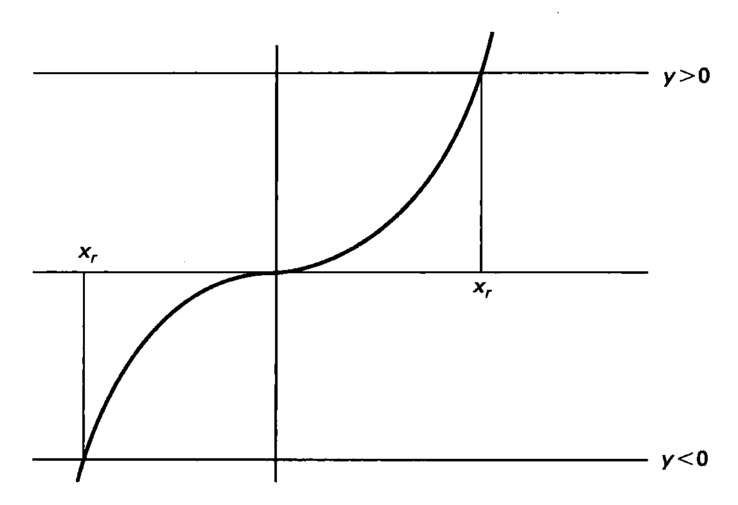


Figure 3.2-17 Roots of $g(x) = x^n - y = 0$ when n is odd.

Example 3.2-10

(*nonlinear devices*) A number of nonlinear zero-memory devices can be modeled by a transmission function $g(x) = x^n$. Let $Y = X^n$. The pdf of Y depends on whether n is even or odd. We solve the case of n odd, leaving n even as an exercise. For n odd and $y > 0$, the only real root to $y - x^n = 0$ is $x_r = y^{1/n}$. Also

$$\frac{dg}{dx} = nx^{n-1} = ny^{(n-1)/n}.$$

For $y < 0$, the only real root is $x_r = -|y|^{1/n}$. See Figure 3.2-17. Also

$$\frac{dg}{dx} = n|y|^{(n-1)/n}.$$

Hence

$$f_Y(y) = \begin{cases} \frac{1}{n} y^{(1-n)/n} \cdot f_X(y^{1/n}), & y \geq 0, \\ \frac{1}{n} |y|^{(1-n)/n} \cdot f_X(-|y|^{1/n}), & y < 0. \end{cases}$$

In problems in which $g(x)$ assumes a constant value, say $g(x) = c$, over some nonzero width interval Equation 3.2-23 cannot be used to compute $f_Y(y)$ because $g'(x) = 0$ over the interval. One additionally has to find the probability mass generated by this flat section.

Example 3.2-11

(*linear amplifier with cutoff.*) Consider a nonlinear device with transformation as shown in Figure 3.2-18.
The function $g(x)$ is given by

$$g(x) = 0, \quad |x| \geq 1 \tag{3.2-24}$$

$$g(x) = x, \quad -1 < x < 1. \tag{3.2-25}$$

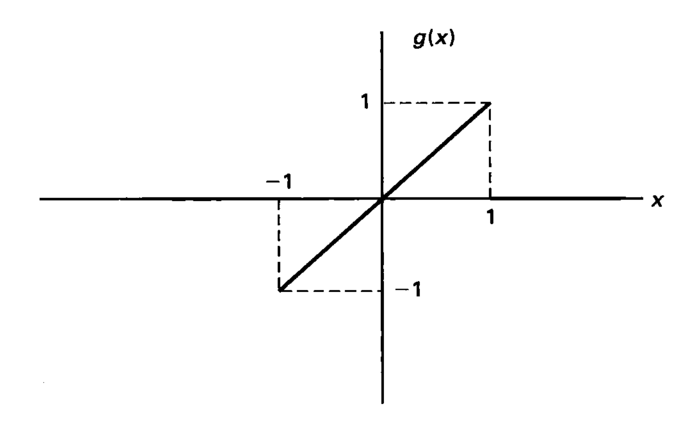


Figure 3.2-18 A linear amplifier with cutoff.

Thus $g'(x) = 0$ for $|x| \geq 1$, and $g'(x) = 1$ for $-1 < x < 1$. For $y \geq 1$ and $y \leq -1$, there are no real roots to $y - g(x) = 0$. Hence $f_Y(y) = 0$ in this range. For $-1 < y < 1$, the only root to $y - g(x) = y - x = 0$ is $x = y$. Hence in this range Equation 3.2-23 applies with $|g'(x)| = 1$ and so $f_Y(y) = f_X(y)$. We note that $P[Y = 0] = P[X \geq 1] + P[X \leq -1]$. If X: $N(0, 1)$, $P[X \geq 1] = 1/2 - \text{erf}(1) = P[X \leq -1]$ and so $P[Y = 0] = 1 - 2\text{erf}(1) = 0.317$. We would like to incorporate the result that $P[Y = 0] = 0.317$ into the pdf of Y. We can do this with the aid of delta functions realizing that

$$P[Y = 0] = 0.317 = \lim_{\epsilon \to 0} \int_{0-\epsilon}^{0+\epsilon} 0.317\delta(y)dy.$$

Hence by including the term $0.317\delta(y)$ in $f_Y(y)$ we obtain the complete solution as:

$$f_Y(y) = \begin{cases} 0, & |y| \geq 1, \\ (2\pi)^{-1/2} \exp\left(-\frac{1}{2}y^2\right) + 0.317\delta(y), & -1 < y < 1. \end{cases}$$

Example 3.2-12

(*infinite roots*) Here we consider the periodic extension of the transformation shown in Figure 3.2-18. The extended $g(x)$ is shown in Figure 3.2-19.

The function in this case is described by

$$g(x) = \sum_{n=-\infty}^{\infty} (x - 2n) \text{ rect}\left(\frac{x - 2n}{2}\right).$$

Here rect is the symmetric unit-pulse function defined as

$$\text{rect}(x) = \begin{cases} 1, & -0.5 < x < +0.5, \\ 0, & \text{else.} \end{cases}$$

As in the previous example $f_Y(y) = 0$ for $|y| \geq 1$ because there are no real roots to the equation $y - g(x) = 0$ in this range. On the other hand, when $-1 < y < 1$, there are an infinite number of roots to $y - g(x) = 0$ and these are given by $x_n = y + 2n$ for

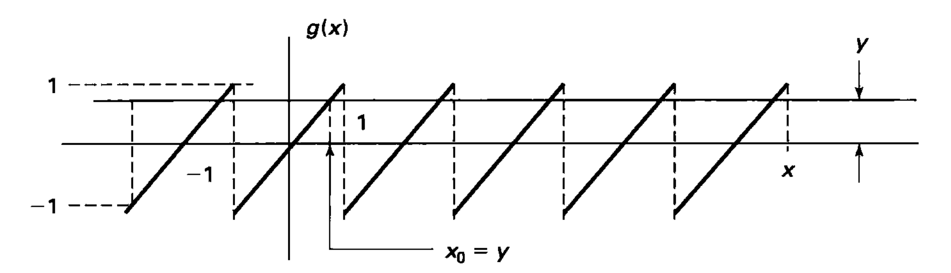


Figure 3.2-19 Periodic transformation function.

$n = \ldots, -2, -1, 0, 1, 2, \ldots$. At each root $|g'(x_n)| = 1$. Hence, from Equation 3.2-23 we obtain $f_Y(y) = \sum_{n=-\infty}^{\infty} f_X(y+2n)\ \text{rect}(\frac{y}{2})$. In the case that X: $N(0,1)$ this specializes to

$$f_Y(y) = (2\pi)^{-1/2} \sum_{n=-\infty}^{\infty} \exp\left(-\frac{1}{2}(y+2n)^2\right) \times \text{rect}\left(\frac{y}{2}\right).$$

While this result is correct, it seems hard to believe that the sum of infinite positive terms yields a function whose area is restricted to one. To show that $f_Y(y)$ does indeed integrate to one, we proceed as follows:

$$\int_{-\infty}^{\infty} f_Y(y)dy = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \int_{-1}^{1} \exp\left(-\frac{1}{2}(y+2n)^2\right) dy \tag{3.2-26}$$

$$= \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{\infty} \int_{-1+2n}^{1+2n} \exp\left(-\frac{1}{2}y^2\right) dy \tag{3.2-27}$$

$$= \sum_{n=-\infty}^{\infty} [\text{erf}(1+2n) - \text{erf}(-1+2n)]. \tag{3.2-28}$$

If this last sum is written out, the reader will quickly find that all the terms cancel except the first ($n = -\infty$) and the last ($n = \infty$). This leaves that

$$\int_{-\infty}^{\infty} f_Y(y)dy = \text{erf}(\infty) - \text{erf}(-\infty) = 2 \times \text{erf}(\infty) = 1.$$

3.3 SOLVING PROBLEMS OF THE TYPE $Z = g(X,Y)$

In many problems in science and engineering, a random variable Z is functionally related to two (or more) random variables X, Y. Some examples are

1. The signal Z at the input of an amplifier consists of a signal X to which is added independent random noise Y. Thus $Z = X + Y$. If X is also an RV, what is the pdf of Z? (See Figure 3.3-1.)

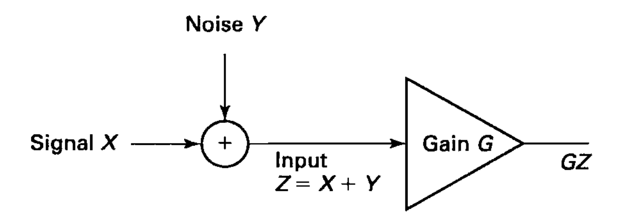


Figure 3.3-1 The signal plus independent additive noise problem.

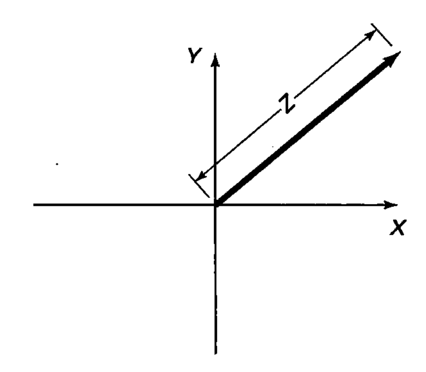


Figure 3.3-2 Displacement in the random-walk problem.

2. A two-engine airplane is capable of flight as long as at least one of its two engines is working. If the time-to-failures of the starboard and port engines are X and Y, respectively, then the time-to-crash of the airplane is $Z \triangleq \max(X, Y)$. What is the pdf of Z?
3. Many signal processing systems multiply two signals together (modulators, demodulators, correlators, and so forth). If X is the signal on one input and Y is the signal on the other input, what is the pdf of the output $Z \triangleq XY$?
4. In the famous "random-walk" problem that applies to a number of important physical problems, a particle undergoes random independent displacements X and Y in the x and y directions, respectively. What is the pdf of the total displacement $Z \triangleq [X^2 + Y^2]^{1/2}$? (See Figure 3.3-2.)

Problems of the type $Z = g(X, Y)$ are not fundamentally different from the type of problem we discussed in Section 3.2. Recall that for $Y = g(X)$ the basic problem was to find the point set C_y such that the events $\{\zeta\colon Y(\zeta) \leq y\}$ and $\{\zeta\colon X(\zeta) \in C_y\}$ were equal. Essentially, the same problem occurs here as well: Find the point set C_z in the (x, y) plane such that the events $\{\zeta\colon Z(\zeta) \leq z\}$ and $\{\zeta\colon X(\zeta), Y(\zeta) \in C_z\}$ are equal, this being indicated in our usual shorthand notation by

$$\{Z \leq z\} = \{(X, Y) \in C_z\} \tag{3.3-1}$$

and

$$F_Z(z) = \iint_{(x,y)\in C_z} f_{XY}(x, y)dx\,dy. \tag{3.3-2}$$

The point set C_z is determined from the functional relation $g(x, y) \leq z$. Clearly in problems of the type $Z = g(X, Y)$ we deal with joint densities or distributions and double integrals (or summations) instead of single ones. Thus, in general, the computation of $f_Z(z)$ is more complicated than the computation of $f_Y(y)$ in $Y = g(X)$. However, we have access to two great labor-saving devices, which we shall learn about later: (1) We can solve many $Z = g(X, Y)$-type problems by a "turn-the-crank" type formula, essentially an extension of Equation 3.2-23, through the use of *auxiliary* variables (Section 3.4); and (2) we can solve problems of the type $Z = X + Y$ through the use of *characteristic functions* (Chapter 4). However, use of these shortcut methods at this stage would obscure the underlying principles.

Let us now solve the problems mentioned earlier from first principles.

Example 3.3-1

(*product of RVs*) To find C_z in Equation 3.3-2 for the CDF of $Z = XY$, we need to determine the region where $g(x, y) \triangleq xy \leq z$. This region is shown in Figure 3.3-3 for $z > 0$.

Thus, reasoning from the diagram, we compute

$$F_Z(z) = \int_0^\infty \left(\int_{-\infty}^{z/y} f_{XY}(x, y) dx \right) dy + \int_{-\infty}^0 \left(\int_{z/y}^\infty f_{XY}(x, y) dx \right) dy \quad \text{for } z \geq 0. \quad (3.3\text{-}3)$$

To compute the density f_Z, it is necessary to differentiate this expression with respect to z. We can do this directly on Equation 3.3-3; however, to see this more clearly we first define the indefinite integral $G_{XY}(x, y)$ by

$$G_{XY}(x, y) \triangleq \int f_{XY}(x, y) dx. \quad (3.3\text{-}4)$$

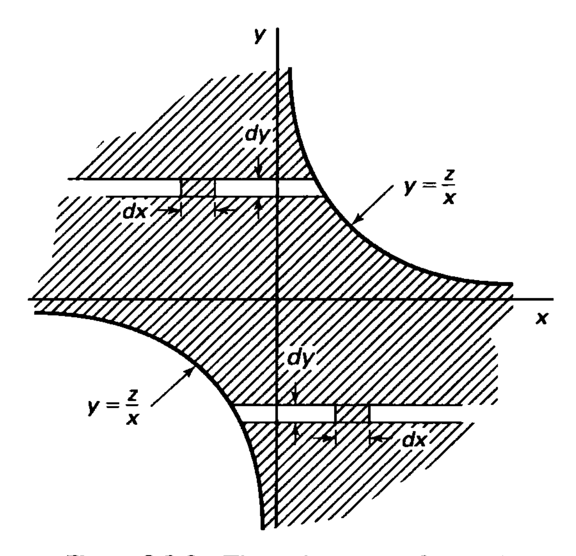


Figure 3.3-3 The region $xy \leq z$ for $z > 0$.

Then

$$F_Z(z) = \int_0^\infty [G_{XY}(z/y, y) - G_{XY}(-\infty, y)]dy$$
$$+ \int_{-\infty}^0 [G_{XY}(\infty, y) - G_{XY}(z/y, y)]dy$$

and differentiation with respect to z is fairly simple now to get

$$f_Z(z) = \frac{dF_Z(z)}{dz}$$
$$= \int_{-\infty}^{\infty} \frac{1}{|y|} f_{XY}(z/y, y)dy. \tag{3.3-5}$$

We could have gotten the same answer by directly differentiating Equation 3.3-3 with respect to z using formula A2-1 of Appendix A.

The question remains as to what is the answer when $z < 0$. It turns out that Equation 3.3-7 is valid for $z < 0$ as well, so that it is valid for all z. The corresponding sketch in the case when $z < 0$ is shown in Figure 3.3-4. From this figure, performing the integration over the new shaded region corresponding to $\{xy \leq z\}$ now in the case $z < 0$, you should get the same integral expression for $F_Z(z)$ as above, that is, Equation 3.3-3. Taking the derivative with respect to z and moving it inside the integral over y, we then again obtain Equation 3.3-7. Thus, the general pdf for the product of two random variables for any value of z is confirmed to be

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f_{XY}(z/y, y)dy, \qquad -\infty < z < +\infty. \tag{3.3-6}$$

As a special case, assume X and Y are independent, identically distributed (i.i.d.) RVs with

$$f_X(x) = f_Y(x) \triangleq \frac{\alpha/\pi}{\alpha^2 + x^2}.$$

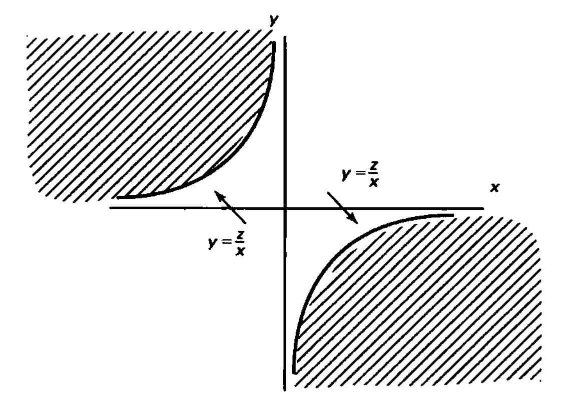


Figure 3.3-4 The region $xy \leq z$ for $z < 0$.

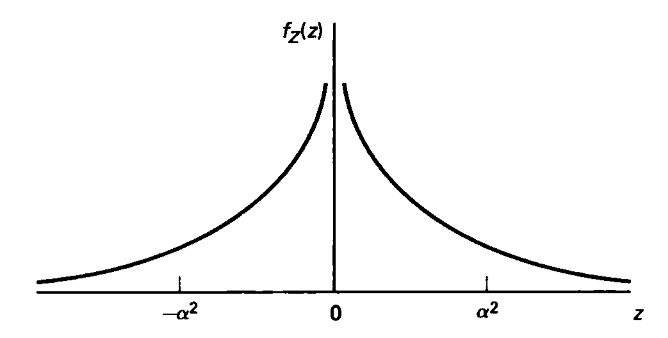


Figure 3.3-5 The pdf $f_Z(z)$ of $Z = XY$ when X and Y are i.i.d. RVs and Cauchy.

This is known as the Cauchy† probability law. Because of independence

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

and because of the evenness of the integrand in Equation 3.3-7, we obtain,‡ after a change of variable,

$$\begin{aligned} f_Z(z) &= \left(\frac{\alpha}{\pi}\right)^2 \int_0^\infty \frac{1}{z^2 + \alpha^2 x} \cdot \frac{1}{\alpha^2 + x} dx \\ &= \left(\frac{\alpha}{\pi}\right)^2 \frac{1}{z^2 - \alpha^4} \ln \frac{z^2}{\alpha^4}. \end{aligned} \tag{3.3-7}$$

See Figure 3.3-5 for a sketch of $f_Z(z)$ for $\alpha = 1$.

Example 3.3-2

(*maximum operation*) We wish to compute the pdf of $Z = \max(X, Y)$ if X and Y are independent RVs. Then

$$F_Z(z) = P[\max(X, Y) \leq z].$$

But the event $\{\max(X, Y) \leq z\}$ is equal to $\{X \leq z, Y \leq z\}$. Hence

$$P[Z \leq z] = P[X \leq z, Y \leq z] = F_X(z) F_Y(z) \tag{3.3-8}$$

and by differentiation, we get

$$f_Z(z) = f_Y(z) F_X(z) + f_X(z) F_Y(z). \tag{3.3-9}$$

Again as a special case, let $f_X(x) = f_Y(x)$ be the uniform $[0, 1]$ pdf. Then

$$f_Z(z) = 2z[u(z) - u(z-1)], \tag{3.3-10}$$

which is ploted in Figure 3.3-6. The computation of $Z = \min(X, Y)$ is left as an end-of-chapter problem.

†Auguste Louis Cauchy (1789–1857). French mathematician who wrote copiously on astronomy, optics, hydrodynamics, function theory, and the like.

‡See B. O. Pierce and R. M. Foster, *A Short Table of Integrals*, 4th ed. (Boston, MA: Ginn & Company, 1956), p. 8.

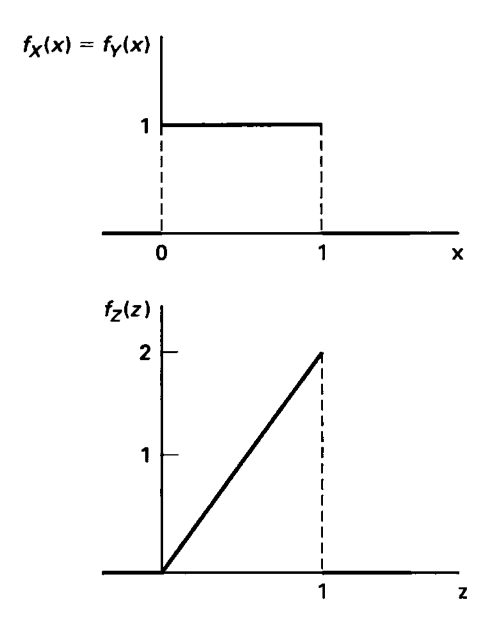


Figure 3.3-6 The pdf of $Z = \max(X, Y)$ for X, Y i.i.d. and uniform in $[0, 1]$.

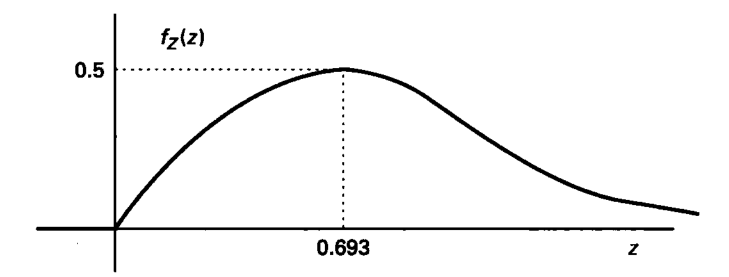


Figure 3.3-7 The pdf of the maximum of two independent exponential random variables.

Example 3.3-3

(*max of exponentials*) Let X, Y be i.i.d. RVs with exponential pdf $f_X(x) = e^{-x}u(x)$. Let $Z = \max(X, Y)$. Compute $f_Z(z)$ and then determine the probability $P[Z \leq 1]$.

Solution From $P[Z \leq z] = P[X \leq z, Y \leq z] = P[X \leq z]P[Y \leq z]$, we obtain

$$F_Z(z) = F_X(z)F_Y(z) = (1 - e^{-z})^2u(z)$$

and

$$f_Z(z) = \frac{dF_Z(z)}{dz} = 2e^{-z}(1 - e^{-z})u(z).$$

The pdf is shown in Figure 3.3-7. Finally, $F_Z(1) = (1 - e^{-1})^2u(1) \approx 0.4$.

The sum of two independent random variables. The situation modeled by $Z = X + Y$ $\left(\text{and its extension } Z = \sum_{i=1}^{N} X_i\right)$ occurs so frequently in engineering and science that the computation of $f_Z(z)$ is perhaps the most important of all problems of the type $Z = g(X, Y)$.

As in other problems of this type, we must find the set of points C_z such that the event $\{Z \leq z\}$ that, by definition, is equal to the event $\{X + Y \leq z\}$ is also equal to $\{(X, Y) \in C_z\}$. The set of points C_z is the set of all points such that $g(x, y) \triangleq x + y \leq z$ and therefore represents the shaded region to the left of the line in Figure 3.3-8; any point in the shaded region satisfies $x + y \leq z$.

Using Equation 3.3-2, specialized for this case, we obtain

$$\begin{aligned} F_Z(z) &= \iint_{x+y\leq z} f_{XY}(x, y)dx\,dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f_{XY}(x, y)dx \right) dy \\ &= \int_{-\infty}^{\infty} [G_{XY}(z - y, y) - G_{XY}(-\infty, y)]dy, \end{aligned} \tag{3.3-11}$$

where $G_{XY}(x, y)$ is the indefinite integral

$$G_{XY}(x, y) \triangleq \int f_{XY}(x, y)dx. \tag{3.3-12}$$

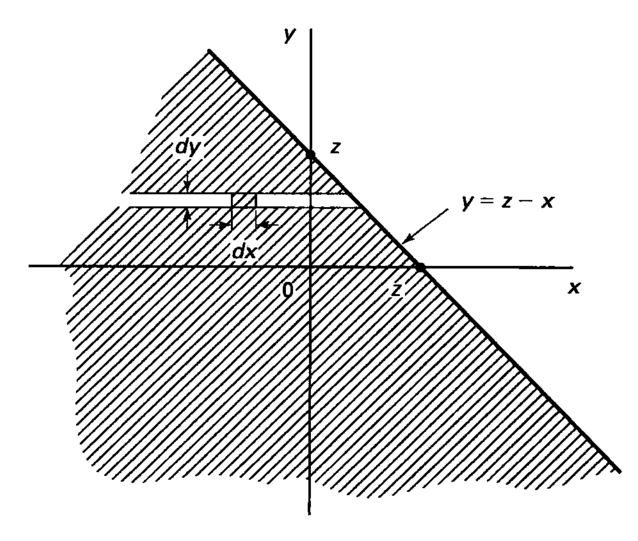


Figure 3.3-8 The region C_z (shaded) for computing the pdf of $Z \triangleq X + Y$.

The pdf is obtained by differentiation of $F_Z(z)$. Thus,

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{\infty} \frac{d}{dz}[G_{XY}(z-y,y)]dy$$

$$= \int_{-\infty}^{\infty} f_{XY}(z-y,y)dy. \tag{3.3-13}$$

Equation 3.3-13 is an important result (compare with Equation 3.3-6 for $Z = XY$). In many instances X and Y are independent RVs so that $f_{XY}(x,y) = f_X(x)f_Y(y)$. Then Equation 3.3-13 takes the special form

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy, \tag{3.3-14}$$

which is known as the *convolution integral* or, more specifically, the *convolution of* f_X *with* f_Y.† It is a simple matter to prove that Equation 3.3-14 can be rewritten as

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx, \tag{3.3-15}$$

by use of the transformation of variables $x = z - y$ in Equation 3.3-14.

Example 3.3-4

(*addition of RVs*) Let X and Y be independent RVs with $f_X(x) = e^{-x}u(x)$ and $f_Y(y) = \frac{1}{2}[u(y+1) - u(y-1)]$ and let $Z \triangleq X + Y$. What is the pdf of Z?

Solution A big help in solving convolution-type problems is to keep track of what is going on graphically. Thus, in Figure 3.3-9(a) is shown $f_X(y)$ and $f_Y(y)$; in Figure 3.3-9(b) is shown $f_X(z-y)$. Note that $f_X(z-y)$ is the *reverse and shifted* image of $f_X(y)$. How do we know that the point at the leading edge of the reverse/shifted image is $y = z$? Consider

$$f_X(z-y) = e^{-(z-y)}u(z-y).$$

But $u(z-y) = 0$ for $y > z$. Therefore the reverse/shifted function is nonzero for $(-\infty, z]$ and the leading edge of $f_X(z-y)$ is at $y = z$.

Since f_X and f_Y are discontinuous functions, we do not expect $f_Z(z)$ to be described by the same expression for all values of z. This means that we must do a careful step-by-step evaluation of Equation 3.3-14 for different regions of z-values.

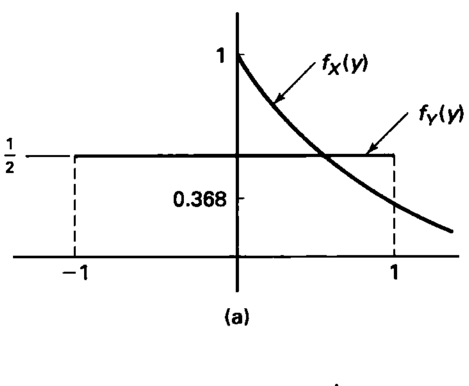
(a) *Region 1.* $z < -1$. For $z < -1$ the situation is as shown in Figure 3.3-10(a). Since there is no overlap, Equation 3.3-14 yields zero. Thus $f_Z(z) = 0$ for $z < -1$.

(b) *Region 2.* $-1 \le z < 1$. In this region the situation is as in Figure 3.3-10(b). Thus Equation 3.3-14 yields

$$f_Z(z) = \frac{1}{2}\int_{-1}^{z} e^{-(z-y)}dy$$

$$= \frac{1}{2}[1 - e^{-(z+1)}].$$

†A common notation for the convolution integral as in Equation 3.3-15 is $f_Z = f_X * f_Y$.



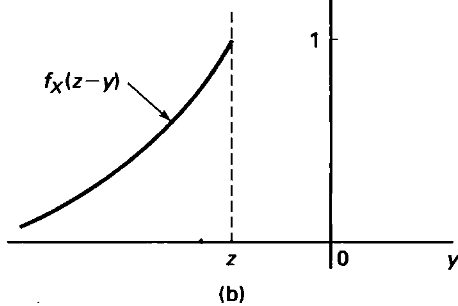


Figure 3.3-9 (a) The pdf's $f_X(y)$, $f_Y(y)$; (b) the reverse/shifted pdf $f_X(z-y)$.

(c) *Region 3.* $z \geq 1$. In this region the situation is as in Figure 3.3-10(c). From Equation 3.3-14 we obtain

$$f_Z(z) = \frac{1}{2}\int_{-1}^{1} e^{-(z-y)}dy$$
$$= \frac{1}{2}[e^{-(z-1)} - e^{-(z+1)}].$$

Before collecting these results to form a graph we make one final important observation: Since no delta functions were involved in the computation, $f_Z(z)$ must be a *continuous* function of z. Hence, as a check on the solution, the $f_Z(z)$ values at the boundaries of the regions must match. For example, at the junction $z = 1$ between region 2 and region 3

$$\tfrac{1}{2}[1 - e^{-(z+1)}]_{z=1} \stackrel{?}{=} \tfrac{1}{2}[e^{-(z-1)} - e^{-(z+1)}]_{z=1}.$$

Obviously the right and left sides of this equation agree so we have some confidence in our solution (Figure 3.3-11).

Equations 3.3-14 and 3.3-15 can easily be extended to computing the pdf of $Z = aX + bY$. To be specific, let $a > 0$, $b > 0$. Then the region $g(x, y) \triangleq ax + by \leq z$ is to the left of the line $y = z/b - ax/b$ (Figure 3.3-12).

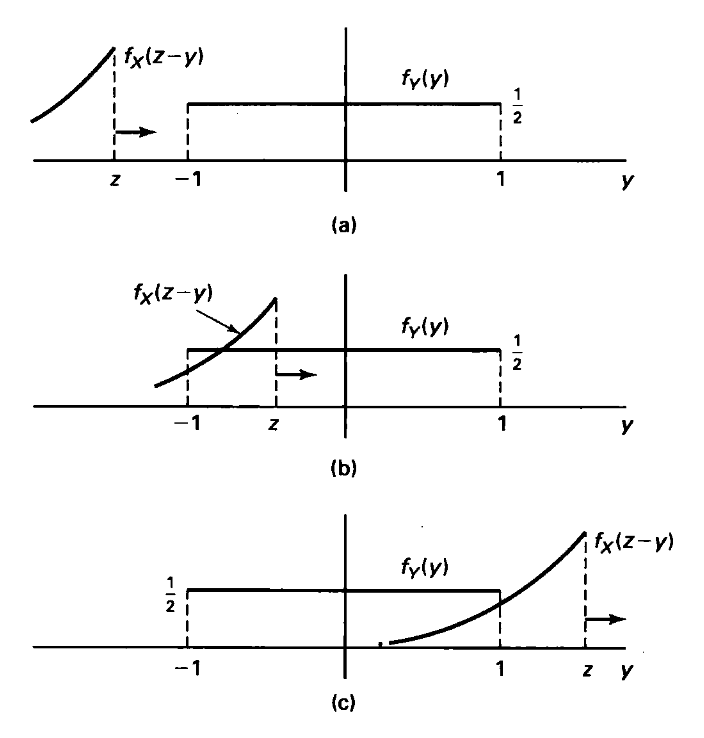


Figure 3.3-10 Relative positions $f_X(z-y)$ and $f_Y(y)$ for (a) $z < 1$; (b) $-1 \leq z < 1$; (c) $z > 1$.

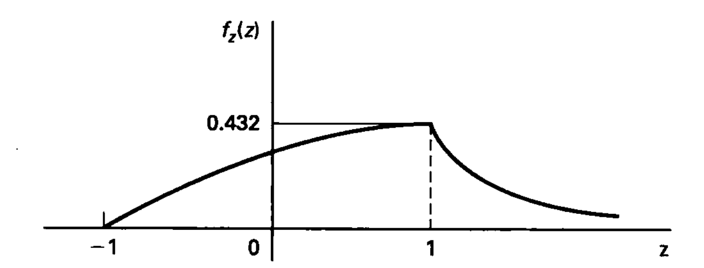


Figure 3.3-11 The pdf $f_Z(z)$ from Example 3.3-4.

Hence

$$
\begin{aligned}
F_Z(z) &= \iint_{g(x,y)\leq z} f_{XY}(x,y)dx\,dy \\
&= \int_{-\infty}^{\infty} f_Y(y)\left(\int_{-\infty}^{z/a-by/a} f_X(x)dx\right)dy. \qquad (3.3\text{-}16)
\end{aligned}
$$

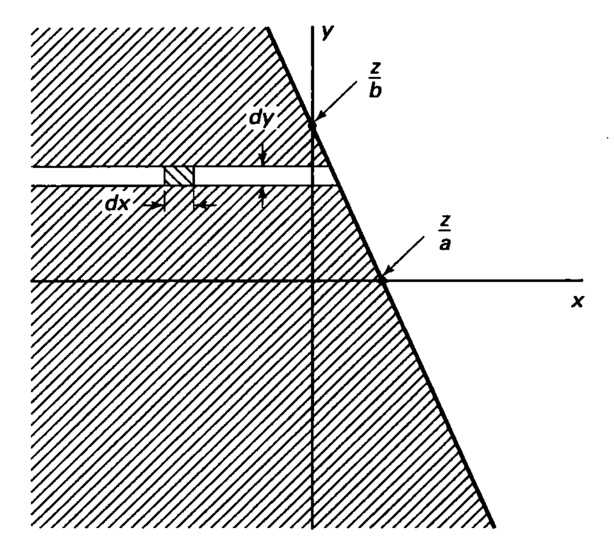


Figure 3.3-12 The region of integration for computing the pdf of $Z = aX + bY$ shown for $a > 0$, $b > 0$.

As usual, to obtain $f_Z(z)$ we differentiate with respect to z; this furnishes

$$f_Z(z) = \frac{1}{a}\int_{-\infty}^{\infty} f_X\left(\frac{z}{a} - \frac{by}{a}\right) f_Y(y)dy, \tag{3.3-17}$$

where we assumed that X and Y are independent RVs. Equivalently, we can compute $f_Z(z)$ by writing

$$V \triangleq aX$$

$$W \triangleq bY$$

$$Z \triangleq V + W.$$

Then again, assuming $a > 0$, $b > 0$ and X, Y independent, we obtain from Equation 3.3-14

$$f_Z(z) = \int_{-\infty}^{\infty} f_V(z - w) f_W(w)dw,$$

where, from Equation 3.2-2,

$$f_V(v) = \frac{1}{a} f_X\left(\frac{v}{a}\right),$$

and

$$f_W(w) = \frac{1}{b} f_Y\left(\frac{w}{b}\right).$$

Thus,

$$f_Z(z) = \frac{1}{ab}\int_{-\infty}^{\infty} f_X\left(\frac{z - w}{a}\right) f_Y\left(\frac{w}{b}\right) dw. \tag{3.3-18}$$

Although Equation 3.3-18 doesn't "look" like Equation 3.3-17, in fact it is identical to it. We need only make the change of variable $y \triangleq w/b$ in Equation 3.3-18 to obtain Equation 3.3-17.

Example 3.3-5

(*a game of Sic bo*) In many jurisdictions in the United States, the taxes and fees from legal gambling parlors are used to finance public education, build roads, etc. Gambling parlors operate to make a profit and set the odds to their advantage. In the popular game of *Sic bo,* the player bets on the outcome of a simultaneous throw of three dice. Many bets are possible, each with a different payoff. Events that are more likely have a smaller payoff, while events that are less likely have a larger payoff. At one large gambling parlor the set odds are the following:

1. Sum of three dice equals 4 or 17 (60 to 1)
2. Sum of three dice equals 5 or 16 (30 to 1);
3. Sum of three dice equals 6 or 15 (17 to 1);
4. Sum of three dice equals 7 or 14 (12 to 1);
5. Sum of three dice equals 8 or 13 (8 to 1);
6. Sum of three dice equals 9 or 10 or 11 or 12 (6 to 1).

For example, 60 to 1 odds means that if the player bets one dollar and the event occurs, he/she gets 60 dollars back minus the dollar ante. It is of interest to calculate the probabilities of the various events.

Solution All the outcomes involve the sum of three i.i.d. random variables. Let X, Y, Z denote the numbers that show up on the three dice, respectively. We can compute the result we need by two successive convolutions. Thus, for the sum on the faces of two dice, the PMF of $X + Y$, $P_{X+Y}(l) \triangleq \sum_{i=1}^{6} P_X(l-i)P_Y(i)$ and the result is shown in Figure 3.3-13. To compute the PMF of the sum of all three RVs $P_{X+Y+Z}(n)$, we perform

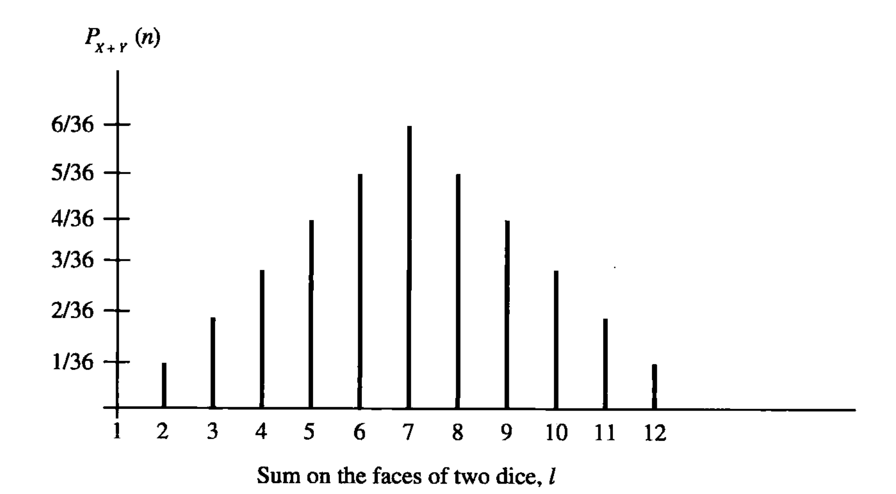


Figure 3.3-13 Probabilities of getting a sum on the faces of two dice.

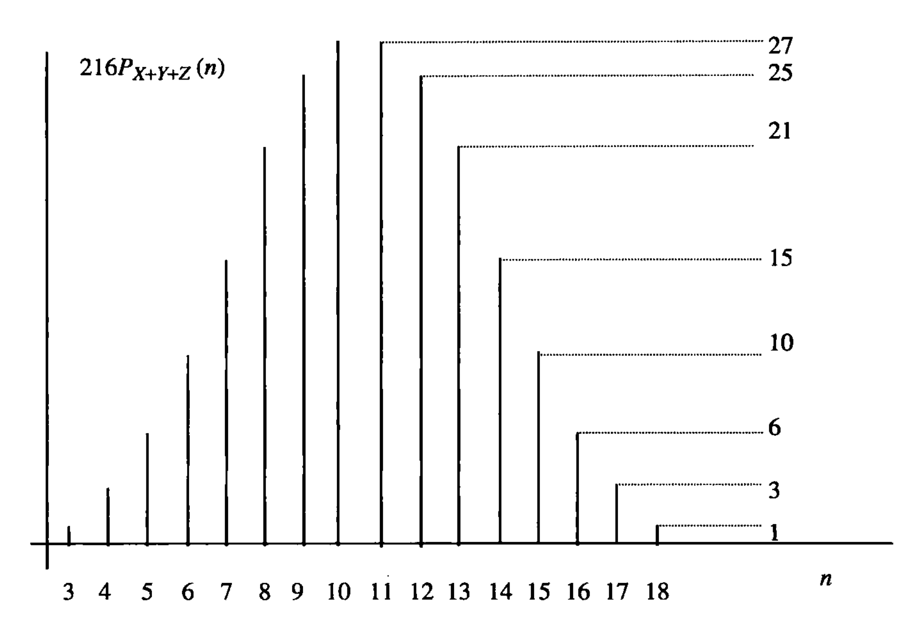


Figure 3.3-14 Probabilities of getting a sum on the faces of three dice.

a second convolution as $P_{X+Y+Z}(n) = \sum_{i=2}^{12} P_Z(n-i)P_{X+Y}(i)$. The result is shown in Figure 3.3-14. From the second convolution, we obtain the probabilities of the events of interest.

We define a "fair payout" (FP) as the return from the house that, on the average yields no loss or gain to the bettor.† If E is the event the bettor bets on, and the ante is \$1.00, then for an FP the return should be 0, so $0 = -\$1.00 + \text{FP} \times P[E]$. So FP $= 1/\text{P[E]}$.

We read the results directly from Figure 3.3-14 to obtain the following:

1. Getting a sum of 4 or 17 (you can bet on either but not both) has a win probability of 3/216 or a fair payout of 72:1 (compare with 60:1).
2. Getting a sum of 5 or 16 (you can bet on either but not both) has a win probability of 6/216 or a fair payout of 36:1 (compare with 30:1).
3. Getting a sum of 6 or 15 (you can bet on either but not both) has a win probability of 10/216 or a fair payout of 22:1 (compare with17:1).
4. Getting a sum of 7 or 14 (you can bet on either but not both) has a win probability of 15/216 or a fair payout of 14:1 (compare with 12:1).
5. Getting a sum of 8 or 13 (you can bet on either but not both) has a win probability of 21/216 or a fair payout of 10:1 (compare with 8:1).

†Obviously the house needs to make enough to cover its expenses for example, salaries, utilities, etc. The definition of a "fair payout" here ignores these niceties. Also the notion of *average* will be explored in some detail in Chapter 4.

6. Getting a sum of 9 or 12 (you can bet on either but not both) has a win probability of 25/216 or a fair payout of 9:1 (compare with 6:1).
7. Getting a sum of 10 or 11 (you can bet on either but not both) has a win probability of 27/216 or a fair payout of 8:1 (compare with 6:1).

Example 3.3-6

(*square-law detector*) Let X and Y be independent RVs, both distributed as $U(-1,1)$. Compute the pdf of $V \triangleq (X+Y)^2$.

Solution We solve this problem in two steps. First, we compute the pdf of $Z \triangleq X+Y$; then we compute the pdf of $V = Z^2$. Using the pulse-width one rect function (see def. on p. 170), we have

$$f_X(x) = \frac{1}{2}\text{rect}\left(\frac{x}{2}\right)$$
$$f_Y(y) = \frac{1}{2}\text{rect}\left(\frac{y}{2}\right)$$
$$f_X(z-y) = \frac{1}{2}\text{rect}\left(\frac{z-y}{2}\right).$$

From Equation 3.3-14 we get

$$f_Z(z) = \frac{1}{4}\int_{-\infty}^{\infty} \text{rect}\left(\frac{y}{2}\right)\text{rect}\left(\frac{z-y}{2}\right)dy. \quad (3.3\text{-}19)$$

The evaluation of Equation 3.3-19 is best done by keeping track graphically of where the "moving," that is, z-dependent function $\text{rect}((z-y)/2)$, is centered vis-à-vis the "stationary," that is, z-independent function $\text{rect}(y/2)$. The term *moving* is used because as z is varied, the function $f_X((z-y)/2)$ has the appearance of moving past $f_Y(y)$. The situation for four different values of z is shown in Figure 3.3-15.

The evaluation of $f_Z(z)$ for the four distinct regions is as follows:

(a) $z < -2$. In this region there is no overlap so

$$f_Z(z) = 0.$$

(b) $-2 \le z < 0$. In this region there is overlap in the interval $(-1, z+1)$ so

$$f_Z(z) = \frac{1}{4}\int_{-1}^{z+1} dy = \frac{1}{4}(z+2).$$

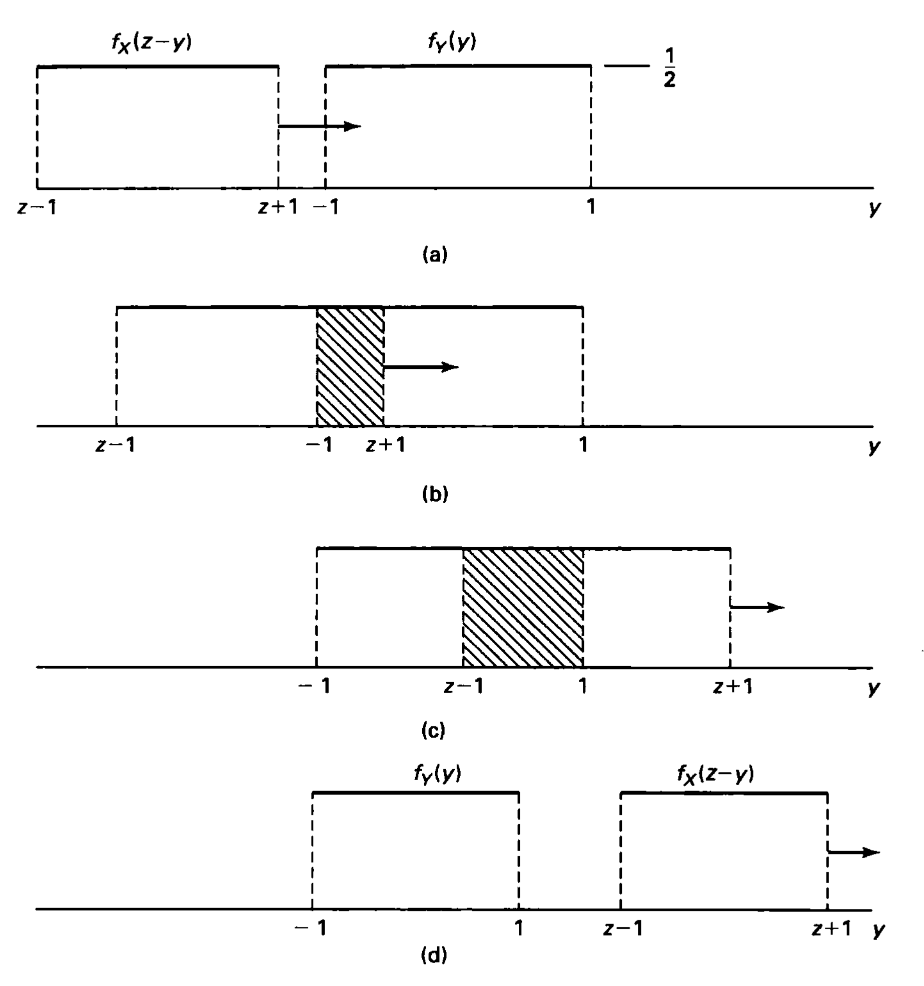


Figure 3.3-15 Four distinct regions in the convolution of two uniform densities: (a) $z < -2$; (b) $-2 \leq z < 0$; (c) $0 \leq z < 2$; (d) $z \geq 2$.

(c) $0 \leq z < 2$. In this region there is overlap in the interval $(z-1, 1)$ so

$$f_Z(z) = \frac{1}{4}\int_{z-1}^{1} dy = \frac{1}{4}(2-z).$$

(d) $2 \leq z$. In this region there is no overlap so

$$f_Z(z) = 0.$$

If we put all these results together, we obtain

$$f_Z(z) = \frac{1}{4}(2-|z|)\text{rect}\left(\frac{z}{4}\right), \tag{3.3-20}$$

which is graphed in Figure 3.3-16.

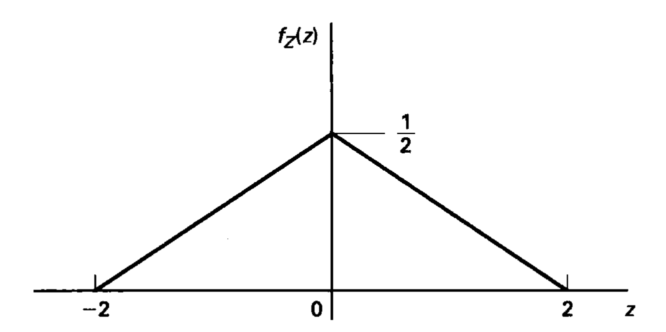


Figure 3.3-16 The pdf of $Z = X + Y$ for X, Y i.i.d. RVs uniform on $(-1, 1)$.

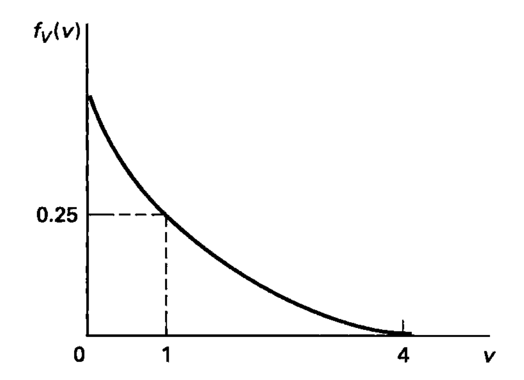


Figure 3.3-17 The pdf of V in Example 3.3-6.

To complete the solution to this problem, we still need the pdf of $V = Z^2$. We compute $f_V(v)$ using Equation 3.3-19 with $g(z) = z^2$. For $v > 0$, the equation $v - z^2 = 0$ has two real roots, that is, $z_1 = \sqrt{v}$, $z_2 = -\sqrt{v}$; for $v < 0$, there are no real roots. Hence, using Equation 3.3-20 in

$$f_V(v) = \sum_{i=1}^{2} f_Z(z_i)/(2|z|)$$

yields

$$f_V(v) = \begin{cases} \frac{1}{4}\left(\frac{2}{\sqrt{v}} - 1\right) & 0 < v \leq 4, \\ 0, & \text{otherwise}, \end{cases} \tag{3.3-21}$$

which is shown in Figure 3.3-17.

The pdf of the sum of discrete random variables can be computed by discrete convolution. For instance, let X and Y be two RVs that take on values $x_1, \ldots, x_k, \ldots$ and $y_1, \ldots, y_j, \ldots$, respectively. Then $Z \triangleq X + Y$ is obviously discrete as well and the PMF is given by

$$P_Z(z_n) = \sum_{x_k+y_j=z_n} P_{X,Y}(x_k, y_j). \tag{3.3-22}$$

If X and Y are independent, Equation 3.3-22 becomes

$$P_Z(z_n) = \sum_{x_k+y_j=z_n} P_X(x_k)P_Y(y_j) = \sum_{x_k} P_X(x_k)P_Y(z_n - x_k). \tag{3.3-23a}$$

If the z_n's and x_k's are equally spaced† then Equation 3.3-23a is recognized as a discrete convolution, in which case it can be written as

$$P_Z(n) = \sum_{all\ k} P_X(k)P_Y(n - k). \tag{3.3-23b}$$

An illustration of the use of Equation 3.3-23b is given below.

Example 3.3-7

(*sum of Bernoulli RVs*) Let B_1 and B_2 be two independent Bernoulli RVs with common PMF

$$P_B(k) = \begin{cases} p, \ k = 1, \\ q, \ k = 0, \\ 0, \ \text{else}, \end{cases} \quad \text{where } q = 1 - p.$$

Let $M \triangleq B_1 + B_2$ and find the PMF $P_M(m)$. We start with the general result

$$P_M(m) = \sum_{b=-\infty}^{+\infty} P_{B_1}(k)P_{B_2}(m - k)$$

$$= \sum_{b=0}^{1} P_{B_1}(k)P_{B_2}(m - k).$$

Since each B_i can only take on values 0 and 1, the allowable values for M are 0, 1, and 2. For all other values of m, $P_M(m) = 0$. This can also be seen graphically from the discrete convolution illustration in Figure 3.3-18.

Calculating the nonzero values of PMF P_M, we obtain

$$P_M(0) = P_{B_1}(0)P_{B_2}(0) = q^2$$

$$P_M(1) = P_{B_1}(0)P_{B_2}(1) + P_{B_1}(1)P_{B_2}(0) = 2pq$$

$$P_M(2) = P_{B_1}(1)P_{B_2}(1) = p^2.$$

The student may notice that M is distributed as binomial $b(k; 2, p)$. Why is this? What would happen if we summed in another independent Bernoulli RV?

†For example, let $z_n = n\Delta$, $x_k = k\Delta$, Δ a constant.

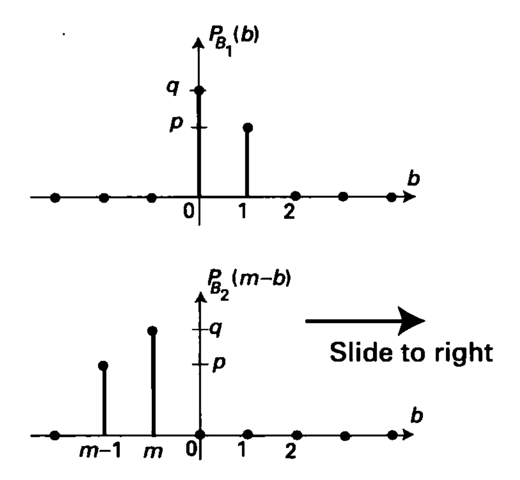


Figure 3.3-18 Illustration of discrete convolution of two Bernoulli PMFs.

Example 3.3-8

(*sum of Poisson RVs*) Let X and Y be two independent Poisson RVs with PMFs $P_X(k) = \frac{1}{k!}e^{-a}a^k$ and $P_Y(i) = \frac{1}{i!}e^{-b}b^i$, where a and b are the Poisson parameters for X and Y, respectively. Let $Z \triangleq X + Y$. Then the PMF of Z, $P_Z(n)$ is given by

$$P_Z(n) = \sum_{k=0}^{n} P_X(k)P_Y(n-k)$$

$$= \sum_{k=0}^{n} \frac{1}{k!}\frac{1}{(n-k)!}e^{-(a+b)}a^k b^{n-k}. \tag{3.3-24}$$

Recall the binomial theorem:

$$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n. \tag{3.3-25}$$

Then

$$P_Z(n) = \frac{1}{n!}e^{-(a+b)}\sum_{k=0}^{n}\binom{n}{k} a^k b^{n-k}$$

$$= \frac{(a+b)^n}{n!}e^{-(a+b)}, \qquad n \geq 0, \tag{3.3-26}$$

which is the Poisson law with parameter $a+b$. Thus, we obtain the important result that the sum of two independent Poisson RVs with parameters a, b is a Poisson RV with parameter $(a+b)$.

Example 3.3-9

(*sum of binomial random variables*) A more challenging example than the previous one involves computing the sums of i.i.d. binomial RVs X and Y. Let $Z = X + Y$; then the PMF $P_Z(m)$ is given by

$$P_Z(m) = \sum_{k=-\infty}^{\infty} P_X(k)P_Y(m-k),$$

where

$$P_X(k) = P_Y(k) = \begin{cases} 0, & k < 0, \\ \binom{n}{k} p^k q^{n-k}, & 0 \le k \le n, \\ 0, & k > n. \end{cases}$$

Thus,

$$\begin{aligned} P_Z(m) &= \sum_{k=\max(0,m-n)}^{\min(n,m)} \binom{n}{k} p^k q^{n-k} \binom{n}{m-k} p^{m-k} q^{n-(m-k)} \\ &= p^m q^{2n-m} \sum_{k=\max(0,m-n)}^{\min(n,m)} \binom{n}{k} \binom{n}{m-k}. \end{aligned}$$

The limits on this summation come from the need to inforce both $0 \le k \le n$ and $0 \le m-k \le n$, the latter being equivalent to $m - n \le k \le m$. Hence the range of the summation must be $\max(0, m-n) \le k \le \min(n, m)$ as indicated.

Somewhat amazingly

$$\sum_{k=\max(0,m-n)}^{\min(n,m)} \binom{n}{k} \binom{n}{m-k} = \binom{2n}{m} \tag{3.3-27}$$

so that we get obtain the PMF of Z as

$$P_Z(m) = \binom{2n}{m} p^m q^{2n-m} \triangleq b(m; 2n, p). \tag{3.3-28}$$

Thus, the sum of two i.i.d. binomial RVs each PMFs $b(k; n, p)$ is a binomial RV with PMF given as $b(k; 2n, p)$.

To show that Equation 3.3-27 is true we first notice that the left-hand side (LHS) has the same value whether $m > n$ (in which case the sum goes from $k = m - n$ up to $k = n$) or whether $m \le n$ (in which case the sum goes from $k = 0$ up to $k = m$). A simple way to see this is to expand out the LHS in both cases. Indeed an expansion of the LHS for $m \le n$ yields

$$\binom{n}{0}\binom{n}{m} + \binom{n}{1}\binom{n}{m-1} + \ldots + \binom{n}{m}\binom{n}{0}. \tag{3.3-29}$$

Doing the expansion in the case $m > n$ to yields the same sum.

Proceeding with the verification of Equation 3.3-27, note that the number of subpopulations of size m that can be formed from a population of size $2n$ is C_m^{2n}. But another way to form these subpopulations is to break the population of size $2n$ into two populations of size n each. Call these two populations of size n each, A and B, respectively. Then the product $C_k^n C_{m-k}^n$ is the number of ways of choosing k subpopulations from A and $m-k$ from B. Then clearly

$$\sum_{k=0}^{m} C_k^n C_{m-k}^n = C_m^{2n} \tag{3.3-30}$$

and since, as we said earlier,

$$\sum_{k=m-n}^{n} C_k^n C_{m-k}^n = \sum_{k=0}^{m} C_k^n C_{m-k}^n,^\dagger$$

the result in Equation 3.3-27 is equally valid when k goes from $m-n$ to n.

In Chapter 4 we will find a simpler method for showing that the sum of i.i.d. binomial RVs is binomial. The method uses transformations called moment generating functions and/or characteristic functions.

We mentioned earlier in Section 3.2 that although the formula in Equation 3.3-23a (and its extensions to be discussed in Section 3.4) is very handy for solving problems of this type, the indirect approach is sometimes easier. We illustrate with the following example.

Example 3.3-10

(*sum of squares*) Let X and Y be i.i.d. RVs with $X{:}N(0,\sigma^2)$. What is the pdf of $Z \triangleq X^2+Y^2$?

Solution We begin with the fundamental result given in Equation 3.3-2:

$$\begin{aligned} F_Z(z) &= \iint_{(x,y)\in C_z} f_{XY}(x,y)\,dx\,dy \quad \text{for} \quad z \geq 0 \\ &= \frac{1}{2\pi\sigma^2}\iint_{x^2+y^2\leq z} e^{-(1/2\sigma^2)(x^2+y^2)}\,dx\,dy. \end{aligned} \tag{3.3-31}$$

The region C_z consists of the shaded region in Figure 3.3-19.

Equation 3.3-31 is easily evaluated using polar coordinates. Let

$$x = r\cos\theta \qquad y = r\sin\theta$$

$$dx\,dy \to r\,dr\,d\theta.$$

---

†This formula can also be verified by using the change of variables $l \triangleq m-k$ in the RHS. The resulting sum will run from large to small, but reversing the summation order does not affect a sum.

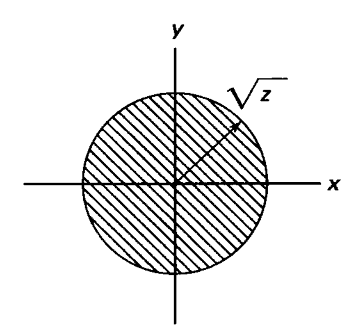


Figure 3.3-19 The region C_z for the event $\{X^2 + Y^2 \leq z\}$ for $z \geq 0$.

Then $x^2 + y^2 \leq z \rightarrow r \leq \sqrt{z}$ and Equation 3.3-31 is transformed into

$$F_Z(z) = \frac{1}{2\pi\sigma^2}\int_0^{2\pi} d\theta \int_0^{\sqrt{z}} r \exp\left(-\frac{1}{2\sigma^2}r^2\right) dr$$

$$= [1 - e^{-z/2\sigma^2}]u(z) \tag{3.3-32}$$

and

$$f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{1}{2\sigma^2}e^{-z/2\sigma^2}u(z). \tag{3.3-33}$$

Thus, $Z = X^2 + Y^2$ is an *exponential* RV if X and Y are i.i.d. zero-mean Gaussian.

Example 3.3-11

(*squareroot of sum of squares*) If the previous example is modified to finding the pdf of $Z \triangleq (X^2 + Y^2)^{1/2}$, a radically different pdf results. Again we use Equation 3.3-2 except that now C_z consists of the shaded region in Figure 3.3-20.

Thus,

$$F_Z(z) = \frac{1}{2\pi\sigma^2}\int_0^{2\pi} d\theta \int_0^{z} r \exp\left(-\frac{1}{2\sigma^2}r^2\right) dr$$

$$= (1 - e^{-z^2/2\sigma^2})u(z) \tag{3.3-34}$$

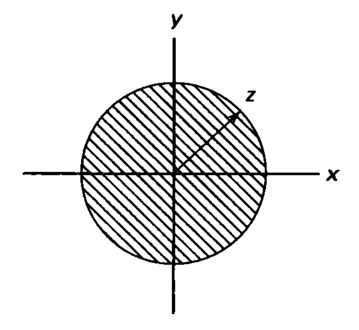


Figure 3.3-20 The region C_z for the event $\{(X^2 + Y^2)^{1/2} \leq z\}$.

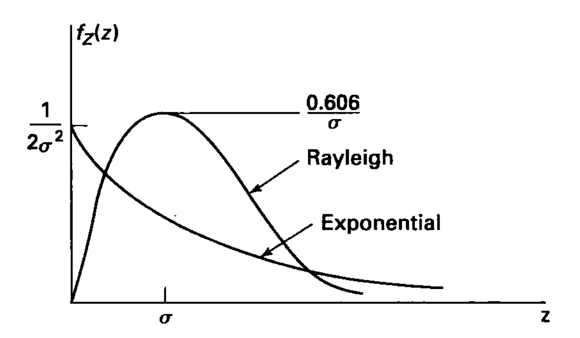


Figure 3.3-21 Rayleigh and exponential pdf's.

and

$$f_Z(z) = \frac{z}{\sigma^2} e^{-z^2/2\sigma^2} u(z), \tag{3.3-35}$$

which is the Rayleigh density function. It is also known as the Chi-square distribution with two degrees of freedom. The exponential and Rayleigh pdf's are compared in Figure 3.3-21.

Stephen O. Rice [3-3], who in the 1940s did pioneering work in the analysis of electrical noise, showed that narrow-band noise signals at center frequency ω can be represented by the wave

$$Z(t) = X \cos \omega t + Y \sin \omega t, \tag{3.3-36}$$

where t is time, ω is the radian frequency in radians per second and where X and Y are i.i.d. RVs distributed as $N(0, \sigma^2)$. The so-called *envelope* $Z \triangleq (X^2 + Y^2)^{1/2}$ has, therefore, a Rayleigh distribution with parameter σ.

The next example generalizes the results of Example 3.3-10 and is a result of considerable interest in communication theory.

***Example 3.3-12**†

(*the Rician density*)‡ S. O. Rice considered a version of the following problem: Let X: $N(P,\ \sigma^2)$ and Y: $N(0, \sigma^2)$ be independent Gaussian RVs. What is the pdf of $Z = (X^2 + Y^2)^{1/2}$? Note that with power parameter $P = 0$, we obtain the solution of Example 3.3-11.

We write

$$F_Z(z) = \begin{cases} \dfrac{1}{2\pi\sigma^2} \displaystyle\iint_{(x^2+y^2)^{1/2} \le z} \exp\left[-\frac{1}{2}\left(\left[\frac{x-P}{\sigma}\right]^2 + \left(\frac{y}{\sigma}\right)^2\right)\right] dx\, dy, & z > 0, \\ 0, \quad z < 0. \end{cases} \tag{3.3-37}$$

---

†Starred examples are somewhat more involved and can be omitted on a first reading.

‡Sometimes called the Rice–Nakagami pdf in recognition of the work of Nakagami around the time of World War II.

The usual Cartesian-to-polar transformation $x = r\cos\theta$, $y = r\sin\theta$, $r = (x^2 + y^2)^{1/2}$, $\theta = \tan^{-1}(y/x)$ yields

$$F_Z(z) = \frac{\exp\left[-\frac{1}{2}\left(\frac{P}{\sigma}\right)^2\right]}{2\pi\sigma^2} \int_0^z e^{-\frac{1}{2}(r/\sigma)^2} \left(\int_0^{2\pi} e^{rP\cos\theta/\sigma^2} d\theta\right) r dr \cdot u(z). \tag{3.3-38}$$

The function

$$I_o(x) \triangleq \frac{1}{2\pi} \int_0^{2\pi} e^{x\cos\theta} d\theta$$

is called the *zero-order modified Bessel function of the first kind* and is monotonically increasing like e^x. With this notation, the cumbersome Equation 3.3-38 can be rewritten as

$$F_Z(z) = \frac{\exp\left[-\frac{1}{2}\left(\frac{P}{\sigma}\right)^2\right]}{\sigma^2} \int_0^z r I_o\left(\frac{rP}{\sigma^2}\right) e^{-\frac{1}{2}(r/\sigma)^2} dr \cdot u(z), \tag{3.3-39}$$

where the step function $u(z)$ ensures that the above is valid for all z. To obtain $f_Z(z)$ we differentiate with respect to z. This produces

$$f_Z(z) = \frac{z}{\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{P^2 + z^2}{\sigma^2}\right)\right] I_o\left(\frac{zP}{\sigma^2}\right) \cdot u(z). \tag{3.3-40}$$

The pdf given in Equation 3.3-40 is called the *Rician* probability density. Since $I_o(0) = 1$, we obtain the Rayleigh law when $P = 0$. When $zP \gg \sigma^2$, that is, the argument of $I_o(\cdot)$ is large, we use the approximation

$$I_o(x) \approx \frac{e^x}{(2\pi x)^{1/2}}$$

to obtain

$$f_Z(z) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \left(\frac{z}{P}\right)^{1/2} e^{-\frac{1}{2}[(z-P)/\sigma]^2},$$

which is almost Gaussian [except for the factor $(z/P)^{1/2}$]. This is the pdf of the *envelope* of the sum of a strong sine wave and weak narrow-band Gaussian noise, a situation that occurs not infrequently in electrical communications.

3.4 SOLVING PROBLEMS OF THE TYPE $V = g(X, Y)$, $W = h(X, Y)$

The problem of two functions of two random variables is essentially an extension of the earlier cases except that the algebra is somewhat more involved.

Fundamental Problem

We are given two RVs X, Y with joint pdf $f_{XY}(x, y)$ and two differentiable functions $g(x, y)$ and $h(x, y)$. Two new random variables are constructed according to $V = g(X, Y)$, $W = h(X, Y)$. How do we compute the joint CDF $F_{VW}(v, w)$ (or joint pdf $f_{VW}(v, w)$) of V and W?

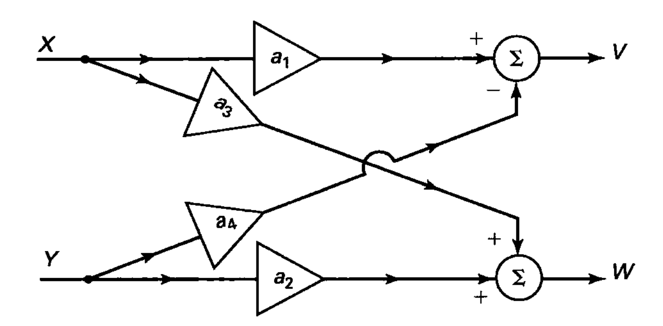


Figure 3.4-1 A two-variable-to-two-variable matrixer.

Illustrations. 1. The transformation shown in Figure 3.4-1 occurs in communication systems such as in the generation of stereo baseband systems [3-2]. The $\{a_i\}$ are gains. When $a_1 = a_2 = \cos\theta$ and $a_3 = a_4 = \sin\theta$, the circuit is known as a *θ-rotational transformer*. In another application if X and Y are used to represent, for example, the left and right pick-up signals in stereo broadcasting, then V and W represent the difference and sum signals if all the a_i's are set to unity. The sum and difference signals are then used to generate the signal to be transmitted. Suppose for the moment that there are no source signals and that X and Y therefore represent only Gaussian noise. What is the pdf of V and W?

2. The error in the landing location of a spacecraft from a prescribed point is denoted by (X, Y) in Cartesian coordinates. We wish to specify the error in polar coordinates $V \triangleq (X^2 + Y^2)^{1/2}$, $W \triangleq \tan^{-1}(Y/X)$. Given the joint pdf $f_{XY}(x, y)$ of landing error coordinates in Cartesian coordinates, how do we compute the pdf of the landing error in polar coordinates?

The solution to the problem at hand is, as before, to find a point set C_{vw} such that the two events $\{V \le v, W \le w\}$ and $\{(X,Y) \in C_{vw}\}$ are equal.† Thus, the fundamental relation is

$$P[V \le v, W \le w] \triangleq F_{VW}(v, w)$$
$$= \iint_{(x,y)\in C_{vw}} f_{XY}(x, y)\,dx\,dy. \tag{3.4-1}$$

The region C_{vw} is given by the points x, y that satisfy

$$C_{vw} = \{(x, y) \colon g(x, y) \le v, h(x, y) \le w\}. \tag{3.4-2}$$

We illustrate the application of Equation 3.4-1 with an example.

Example 3.4-1

(*sum and difference*) We are given $V \triangleq X + Y$ and $W \triangleq X - Y$ and wish to calculate the pdf $f_{VW}(v, w)$. The point set C_{vw} is described by the combined constraints $g(x, y) \triangleq x + y \le v$ and $h(x, y) \triangleq x - y \le w$; it is shown in Figure 3.4-2 for $v > 0$, $w > 0$.

†In more elaborate notation, we would write $\{\zeta \colon V(\zeta) \le v \text{ and } W(\zeta) \le w\} = \{\zeta \colon (X(\zeta), Y(\zeta)) \in C_{vw}\}$.

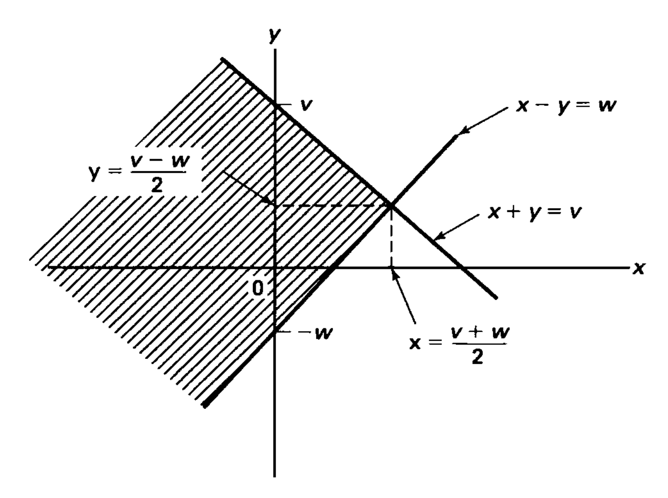


Figure 3.4-2 Point set C_{vw} (shaded region) for Example 3.4-1.

The integration over the shaded region yields

$$F_{VW}(v, w) = \int_{-\infty}^{(v+w)/2} \left(\int_{x-w}^{v-x} f_{XY}(x, y)dy \right) dx. \tag{3.4-3}$$

To obtain the joint density $f_{VW}(v, w)$, we use Equation 2.6-30. Hence

$$\begin{aligned}
f_{VW}(v, w) &= \frac{\partial^2 F_{VW}(v, w)}{\partial v\, \partial w} \\
&= \frac{\partial^2}{\partial v\, \partial w} \int_{-\infty}^{(v+w)/2} \left(\int_{x-w}^{v-x} f_{XY}(x, y)dy \right) dx \\
&= \frac{\partial}{\partial v} \left[\frac{\partial}{\partial w} \int_{-\infty}^{(v+w)/2} \left(\int_{x-w}^{v-x} f_{XY}(x, y)dy \right) dx \right] \\
&= \frac{\partial}{\partial v} \left[\frac{1}{2} \int_{(v-w)/2}^{(v-w)/2} f_{XY}(\frac{v+w}{2}, y)dy + \int_{-\infty}^{(v+w)/2} \left(\frac{\partial}{\partial w} \int_{x-w}^{v-x} f_{XY}(x, y)dy \right) dx \right] \\
&= \frac{\partial}{\partial v} \left[\int_{-\infty}^{(v+w)/2} f_{XY}(x, x - w)dx \right]
\end{aligned}$$

because the first integral is zero for continuous RVs X and Y,

$$\begin{aligned}
&= \frac{\partial}{\partial v} \int_{-\infty}^{(v+w)/2} f_{XY}(x, x - w)dx \\
&= \frac{1}{2} f_{XY} \left(\frac{v+w}{2}, \frac{v-w}{2} \right),
\end{aligned} \tag{3.4-4}$$

where use has been made of the general formula for differentiation of an integral (see Appendix A.2.). Thus, even this simple problem, involving linear functions and just two RVs, requires a considerable amount of work and care to obtain the joint pdf solution. For this reason, problems of the type discussed in this section and their extensions to n RVs, that is, $Y_1 = g_1(X_1, \ldots, X_n)$, $Y_2 = g_2(X_1, \ldots, X_n), \ldots,$ $Y_n = g_n(X_1, \ldots, X_n)$, are generally solved by the technique discussed next called the *direct method* for joint pdf evaluation. Essentially it is the two-dimensional extension of Equation 3.2-23.

Obtaining f_{VW} Directly from f_{XY}

Instead of attempting to find $f_{VW}(v, w)$ through Equation 3.4-1, we can instead take a different approach. Consider the elementary event

$$\{v < V \leq v + dv, w < W \leq w + dw\}$$

and the one-to-one[†] differentiable functions $v = g(x, y)$, $w = h(x, y)$. The inverse mappings exist and are given by $x = \phi(v, w)$, $y = \psi(v, w)$. Later we shall consider the more general case where, possibly, more than one pair of (x_i, y_i) produce a given (v, w).

The probability $P[v < V \leq v + dv, w < W \leq w + dw]$ is the probability that V and W lie in an infinitesimal rectangle of area $dv\,dw$ with vertices at (v, w), $(v + dv, w)$, $(v, w + dw)$, and $(v + dv, w + dw)$. The *image* of this square in the x, y coordinate system is[‡] an infinitesimal parallelogram with vertices at

$$P_1 = (x, y),$$

$$P_2 = \left(x + \frac{\partial \phi}{\partial v} dv, y + \frac{\partial \psi}{\partial v} dv\right),$$

$$P_3 = \left(x + \frac{\partial \phi}{\partial w} dw, y + \frac{\partial \psi}{\partial w} dw\right),$$

$$P_4 = \left(x + \frac{\partial \phi}{\partial v} dv + \frac{\partial \phi}{\partial w} dw, y + \frac{\partial \psi}{\partial v} dv + \frac{\partial \psi}{\partial w} dw\right).$$

This mapping is shown in Figure 3.4-3.

With $\mathscr{R}$ denoting the rectangular region shown in Figure 3.4-3(a) and $\mathscr{S}$ denoting the parallelogram in Figure 3.4-3(b) and $A(\mathscr{R})$ and $A(\mathscr{S})$ denoting the areas of $\mathscr{R}$ and $\mathscr{S}$ respectively, we obtain

$$P[v < V \leq v + dv, w < W \leq w + dw] = \iint_{\mathscr{R}} f_{VW}(\xi, \eta) d\xi\, d\eta \tag{3.4-5}$$

$$= f_{VW}(v, w) A(\mathscr{R}) \tag{3.4-6}$$

$$= \iint_{\mathscr{S}} f_{XY}(\xi, \eta) d\xi\, d\eta \tag{3.4-7}$$

$$= f_{XY}(x, y) A(\mathscr{S}). \tag{3.4-8}$$

[†]Every point (x, y) maps into a unique (v, w) and vice versa.

[‡]See for example [3-4, p.769]

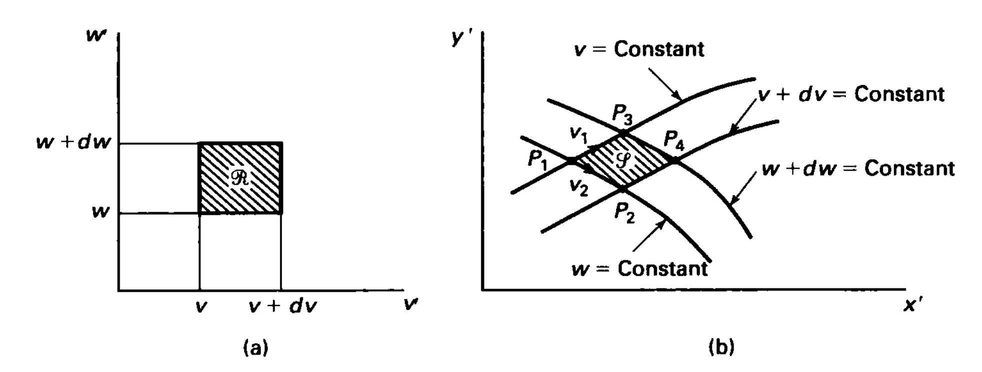


Figure 3.4-3 An infinitesimal rectangle in the v, w system (a) maps into an infinitesimal parallelogram (b) in the x, y system.

Equation 3.4-5 follows from the fundamental relation given in Equation 3.4-1; Equation 3.4-6 follows from the interpretation of the pdf given in Equation 2.4-6; Equation 3.4-7 follows by definition of the point set $\mathscr{S}$, that is, $\mathscr{S}$ is the set of points that makes the events $\{(V, W) \in \mathscr{R}\}$ and $\{(X, Y) \in \mathscr{S}\}$ equal; and Equation 3.4-8 again follows from the interpretation of pdf.

From Equations 3.4-6 and 3.4-8, we find that

$$f_{VW}(v, w) = \frac{A(\mathscr{S})}{A(\mathscr{R})} f_{XY}(x, y), \tag{3.4-9}$$

where $x = \phi(v, w)$ and $y = \psi(v, w)$.

Essentially then, all that remains is to compute the ratio of the two areas. This is done in Appendix C. There we show that the ratio $A(\mathscr{S})/A(\mathscr{R})$ is the magnitude of a quantity called the Jacobian of the transformation $x = \phi(v, w)$, $y = \psi(v, w)$ and given the symbol $\tilde{J}$. If there is more than one solution to the equations $v = g(x, y)$, $w = h(x, y)$, say, $x_1 = \phi_1(v, w)$, $y_1 = \psi_1(v, w)$, $x_2 = \phi_2(v, w)$, $y_2 = \psi_2(v, w), \ldots, x_n = \phi_n(v, w)$, $y_n = \psi_n(v, w)$, then $\mathscr{R}$ maps into multiple, disjoint infinitesimal regions $\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_n$ and $A(\mathscr{S}_i)/A(\mathscr{R}) = |\tilde{J}_i|$, $i = 1, \ldots, n$. The $|\tilde{J}_i|$ are often written as the magnitude of determinants, that is,

$$|\tilde{J}_i| = \text{mag} \begin{vmatrix} \partial\phi_i/\partial v & \partial\phi_i/\partial w \\ \partial\psi_i/\partial v & \partial\psi_i/\partial w \end{vmatrix} = |\partial\phi_i/\partial v \times \partial\psi_i/\partial w - \partial\psi_i/\partial v \times \partial\phi_i/\partial w|. \tag{3.4-10}$$

The end result is the important formula

$$f_{VW}(v, w) = \sum_{i=1}^{n} f_{XY}(x_i, y_i)|\tilde{J}_i|. \tag{3.4-11}$$

It is shown in Appendix C that $|\tilde{J}_i^{-1}| = |J_i| \triangleq |\partial g/\partial x_i \times \partial h/\partial y_i - \partial g/\partial y_i \times \partial h/\partial x_i|$. Then we get the equally important formula

$$f_{VW}(v, w) = \sum_{i=1}^{n} f_{XY}(x_i, y_i)/|J_i|. \tag{3.4-12}$$

Example 3.4-2

(*linear functions*) We are given two functions

$$v \triangleq g(x,y) = 3x + 5y$$

$$w \triangleq h(x,y) = x + 2y \tag{3.4-13}$$

and the joint pdf f_{XY} of two RVs X, Y. What is the joint pdf of two new random variables $V = g(X,Y)$, $W = h(X,Y)$?

Solution The inverse mappings are computed from Equation 3.4-13 to be

$$x = \phi(v,w) = 2v - 5w$$

$$y = \psi(v,w) = -v + 3w.$$

Then

$$\frac{\partial \phi}{\partial v} = 2, \frac{\partial \phi}{\partial w} = -5, \frac{\partial \psi}{\partial v} = -1, \frac{\partial \psi}{\partial w} = 3$$

and

$$|\tilde{J}| = \text{mag}\begin{vmatrix} 2 & -5 \\ -1 & 3 \end{vmatrix} = 1.$$

Assume $f_{XY}(x,y) = (2\pi)^{-1}\exp[-\frac{1}{2}(x^2+y^2)]$. Then, from Equation 3.4-11

$$f_{VW}(v,w) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}[(2v-5w)^2 + (-v+3w)^2]\right]$$

$$= \frac{1}{2\pi}\exp\left[-\frac{1}{2}(5v^2 - 26vw + 34w^2)\right].$$

Thus, the transformation converts uncorrelated Gaussian RVs into correlated Gaussian RVs.

Example 3.4-3

(*two ordered random variables*) Consider two i.i.d., continuous random variables with pdf's $f_{X_1}(x) = f_{X_2}(x) = f_X(x)$. We define two new random variables as $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$. Clearly $Y_1 < Y_2$[†] meaning that realizations of Y_1 are always less than realizations of Y_2. We seek the joint pdf, $f_{Y_1Y_2}(y_1, y_2)$, of Y_1, Y_2 given that

$$Y_1 = g(X_1, X_2) = \min(X_1, X_2)$$

$$Y_2 = h(X_1, X_2) = \max(X_1, X_2).$$

Solution From Figure 3.4-4 (only the first quadrant is shown for convenience but all four quadrants must be considered in any calculation), we see that there are two disjoint real-number, regions and hence two solutions. We note that in $\mathscr{R}_1$, $x_1 > x_2$ while in $\mathscr{R}_2$,

[†]We ignore the zero-probability event $Y_1 = Y_2$.

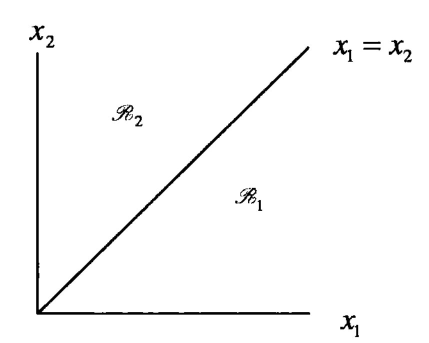


Figure 3.4-4 Showing the two regions of interest for Example 3.3-6.

$x_1 < x_2$. Thus, in $\mathscr{R}_1$ we have $y_1 = x_2, y_2 = x_1$ or, in the g, h notation, $y_1 = g_1(x_1, x_2) = x_2, y_2 = h_1(x_1, x_2) = x_1$. The Jacobian magnitude of this transformation is unity so that $f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}(y_2, y_1) = f_X(y_2)f_X(y_1), y_1 < y_2$.

Repeating this analysis for $\mathscr{R}_2$, we have $y_1 = g_2(x_1, x_2) = x_1, y_2 = h_2(x_1, x_2) = x_2$ and once again the Jacobian magnitude of this transformation of unity. Hence $f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}(y_1, y_2) = f_X(y_1)f_X(y_2), y_1 < y_2$. As always we sum the solutions over the different roots/regions (here there are two) and obtain

$$f_{Y_1Y_2}(y_1, y_2) = \begin{cases} 2f_X(y_1)f_X(y_2), -\infty < y_1 < y_2 < \infty, \\ 0, \text{ else.} \end{cases}$$

Question for the reader: We know that X_1 and X_2 are independent; are Y_1 and Y_2 independent?

Example 3.4-4

(*marginal probabilities of ordered random variables*) In the previous example we ordered two i.i.d. RVs X_1, X_2 as Y_1, Y_2, where $Y_1 < Y_2$. The joint pdf Y_1, Y_2 was shown to be $f_{Y_1Y_2}(y_1, y_2) = 2f_X(y_1)f_X(y_2), -\infty < y_1 < y_2 < \infty$. Here we wish to obtain the marginal pdf's of Y_1, Y_2.

Solution

To get $f_{Y_1}(y_1)$ we have to integrate out $f_{Y_1Y_2}(y_1, y_2) = 2f_X(y_1)f_X(y_2)$, over all $y_2 > y_1$. Hence

$$f_{Y_1}(y_1) = 2f_X(y_1)\int_{y_1}^{\infty} f_X(y_2)dy_2 = 2f_X(y_1)\left(1 - F_X(y_1)\right), \quad -\infty < y_1 < \infty.$$

Likewise, to get $f_{Y_2}(y_2)$ we integrate out $f_{Y_1Y_2}(y_1, y_2) = 2f_X(y_1)f_X(y_2)$, over all $y_1 < y_2$. The result is

$$f_{Y_2}(y_2) = 2f_X(y_2)\int_{-\infty}^{y_2} f_X(y_1)dy_1 = 2f_X(y_2)F_X(y_2), \quad -\infty < y_2 < \infty.$$

Example 3.4-5

(*the minimum and maximum of two Normal random variables*) We wish to see what the pdfs of ordered Normal RVs look like. To that end let X_1, X_2 be i.i.d. Normal $N(0,1)$ RVs pdfs and define $Y_1 \triangleq \min(X_1, X_2)$ and $Y_2 \triangleq \max(X_1, X_2)$. Using the results of Example 3.4-5 we graph these pdf's together with the Normal pdf on the same axes. The curves in Figure 3.4-5 were obtained using the program Microsoft Excel†. The reader may want to duplicate these curves.

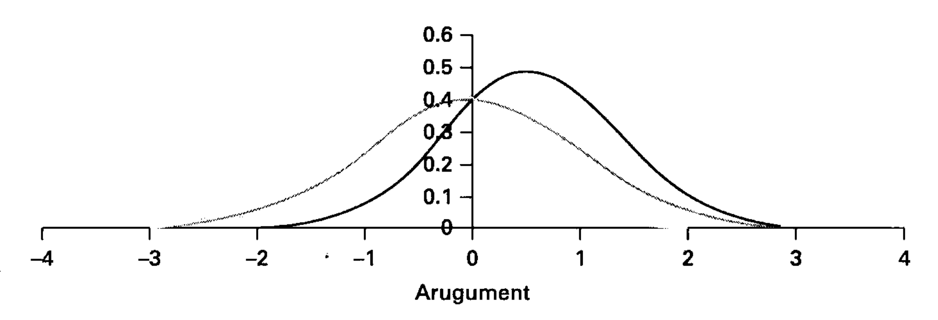


Figure 3.4-5 The pdf of $\min(X_1, X_2)$ peaks at the left of the origin at -0.5 while the pdf of $\max(X_1, X_2)$ peaks at the right of the origin at 0.5. Note that $\mathrm{Var}[\min(X_1, X_2)] = \mathrm{Var}[\min(X_1, X_2)] < 1$.

3.5 ADDITIONAL EXAMPLES

To enable the reader to become familiar with the methods discussed in Section 3.4, we present here a number of additional examples.

Example 3.5-1

(*magnitude and angle*) Consider the RVs

$$V \triangleq g(X,Y) = \sqrt{X^2 + Y^2} \tag{3.5-1}$$

$$W = h(X,Y) = \begin{cases} \tan^{-1}\left(\dfrac{Y}{X}\right), & X > 0, \\ \tan^{-1}\left(\dfrac{Y}{X}\right) + \pi, & X < 0. \end{cases} \tag{3.5-2a}$$

The RV V is called the magnitude or envelope while W is called the phase. Equation 3.5-2a has been written in this form because we seek a solution for w over a 2π interval and the inverse function $\tan^{-1}(y/x)$ has range $(-\pi/2, \pi/2)$ (i.e., its principle value).

†Excel is available with Microsoft Office. The instruction to use Excel are available with the program.

To find the roots of

$$v = \sqrt{x^2+y^2},\ w = \begin{cases} \tan^{-1}\left(\frac{y}{x}\right), & x \geq 0, \\ \tan^{-1}\left(\frac{y}{x}\right) + \pi, & x < 0, \end{cases} \tag{3.5-2b}$$

we observe that for $x \geq 0$, we have $-\frac{\pi}{2} \leq w \leq \frac{\pi}{2}$ and $\cos w \geq 0$. Similarly, for $x < 0$, $\frac{\pi}{2} < w < \frac{3\pi}{2}$ and $\cos w < 0$. Hence the only solution to Equation 3.5-2b is

$$x = v\cos w \triangleq \phi(v,w)$$

$$y = v\sin w \triangleq \psi(v,w).$$

The Jacobian $\tilde{J}$ is given by

$$\tilde{J} = \frac{\partial(\phi,\psi)}{(v,w)} = \begin{vmatrix} \cos w & -v\sin w \\ \sin w & v\cos w \end{vmatrix} = v.$$

Hence the solution is, from Equation 3.4-11,

$$f_{VW}(v,w) = v f_{XY}(v\cos w, v\sin w). \tag{3.5-3}$$

Suppose that X and Y are i.i.d. and distributed as $N(0,\sigma^2)$, that is,

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2}e^{-[(x^2+y^2)/2\sigma^2]}.$$

Then from Equation 3.5-3

$$f_{VW}(v,w) = \begin{cases} \left(\frac{v}{\sigma^2}e^{-v^2/2\sigma^2}\right)\frac{1}{2\pi}, & v > 0, -\frac{\pi}{2} \leq w < \frac{3\pi}{2} \\ 0, & \text{otherwise} \end{cases} \tag{3.5-4}$$

$$= f_V(v)f_W(w).$$

Thus, V and W are independent random variables. The envelope V has a Rayleigh pdf and the phase W is uniform over a 2π interval.

Example 3.5-2

(*magnitude and ratio*) Consider now a modification of the previous problem. Let $V \triangleq \sqrt{X^2+Y^2}$ and $W \triangleq Y/X$. Then with $g(x,y) = \sqrt{x^2+y^2}$ and $h(x,y) = y/x$, the equations

$$v - g(x,y) = 0$$

$$w - h(x,y) = 0$$

have two solutions:

$$x_1 = v(1+w^2)^{-1/2}, \qquad y_1 = wx_1$$

$$x_2 = -v(1+w^2)^{-1/2}, \qquad y_2 = wx_2$$

for $-\infty < w < \infty$ and $v > 0$, and no real solutions for $v < 0$.

A direct evaluation yields $|J_1| = |J_2| = (1+w^2)/v$. Hence

$$f_{VW}(v,w) = \frac{v}{1+w^2}[f_{XY}(x_1,y_1) + f_{XY}(x_2,y_2)].$$

With $f_{XY}(x,y)$ given by

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2}\exp[-(x^2+y^2)/2\sigma^2],$$

we obtain

$$\begin{aligned} f_{VW}(v,w) &= \frac{v}{\sigma^2}e^{-v^2/2\sigma^2}u(v)\cdot\frac{1/\pi}{1+w^2} \\ &= f_V(v)f_W(w). \end{aligned}$$

Thus, the random variables V, W are independent, with V Rayleigh distributed as in Example 3.5-1, and W Cauchy distributed.

Example 3.5-3

(*rotation of coordinates*) Let θ be a prescribed angle and consider the rotational transformation

$$\begin{aligned} V &\triangleq X\cos\theta + Y\sin\theta \\ W &\triangleq X\sin\theta - Y\cos\theta \end{aligned} \tag{3.5-5}$$

with X and Y i.i.d. Gaussian,

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2}e^{-[(x^2+y^2)/2\sigma^2]}.$$

The only solution to

$$\begin{aligned} v &= x\cos\theta + y\sin\theta \\ w &= x\sin\theta - y\cos\theta \end{aligned}$$

is

$$\begin{aligned} x &= v\cos\theta + w\sin\theta \\ y &= v\sin\theta - w\cos\theta. \end{aligned}$$

The Jacobian $\tilde{J}$ is

$$\begin{vmatrix} \dfrac{\partial x}{\partial v} & \dfrac{\partial x}{\partial w} \\ \dfrac{\partial y}{\partial v} & \dfrac{\partial y}{\partial w} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{vmatrix} = -1.$$

Hence

$$f_{VW}(v,w) = \frac{1}{2\pi\sigma^2}e^{-[(v^2+w^2)/2\sigma^2]}.$$

Thus, under the rotational transformation $V = g(X,Y)$, $W = h(X,Y)$ given in Equation 3.5-5, V and W are i.i.d. Gaussian RVs just like X and Y. If X and Y are Gaussian but not independent RVs, it is still possible to find a transformation† so that V, W will be independent Gaussians if the joint pdf of X, Y is Gaussian (Normal).

Example 3.5-4

Consider again the problem of solving for the pdf of $Z = \sqrt{X^2+Y^2}$ as in Example 3.3-11. This time we shall use Equation 3.4-11 to somewhat indirectly compute $f_Z(z)$. First we note that $Z = \sqrt{X^2+Y^2}$ is one function of two RVs while Equation 3.4-11 applies to two functions of two RVs. To convert from one kind of problem to the other, we introduce an *auxiliary variable* $W \triangleq X$. Then

$$Z \triangleq g(X,Y) = \sqrt{X^2+Y^2}$$

$$W \triangleq h(X,Y) = X.$$

The equations

$$z - g(x,y) = 0$$

$$w - h(x,y) = 0$$

have two real roots for $|w| < z$, namely

$$x_1 = w \qquad x_2 = w$$

$$y_1 = \sqrt{z^2-w^2} \qquad y_2 = -\sqrt{z^2-w^2}.$$

At both roots, $|\tilde{J}|$ has the same value:

$$|\tilde{J}_1| = |\tilde{J}_2| = \frac{z}{\sqrt{z^2-w^2}}.$$

Hence a direct application of Equation 3.4-11 yields

$$f_{ZW}(z,w)\frac{z}{\sqrt{z^2-w^2}}[f_{XY}(x_1,y_1) + f_{XY}(x_2,y_2)].$$

Now assume that

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2}e^{-[(x^2+y^2)/2\sigma^2]}.$$

Then, since in this case $f_{XY}(x,y) = f_{XY}(x,-y)$, we obtain

$$f_{ZW}(z,w) = \begin{cases} \dfrac{1}{\pi\sigma^2}\dfrac{z}{\sqrt{z^2-w^2}}e^{-z^2/2\sigma^2}, & z > 0, |w| < z, \\ 0, & \text{otherwise.} \end{cases}$$

†See Chapter 5 on random vectors.

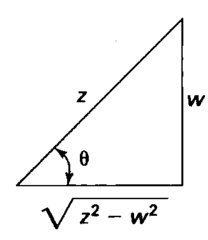


Figure 3.5-1 Trigonometric transformation $w = z\sin\theta$.

However, we don't really want $f_{ZW}(z, w)$, but only the marginal pdf $f_Z(z)$. To obtain this, we use Equation 2.6-47 (with x replaced by z and y replaced by w). This gives

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{ZW}(z, w)dw \\ &= \frac{z}{\sigma^2}e^{-z^2/2\sigma^2}\left[\frac{2}{\pi}\int_0^z \frac{dw}{\sqrt{z^2 - w^2}}\right]u(z). \end{aligned}$$

The term in parentheses has value unity. To see this consider the triangle in Figure 3.5-1 and let $w \triangleq z\sin\theta$. Then $dw = z\cos\theta d\theta$ and $[z^2 - w^2]^{1/2} = z\cos\theta$ and the term in parentheses becomes

$$\frac{2}{\pi}\int_0^z \frac{dw}{\sqrt{z^2 - w^2}} = \frac{2}{\pi}\int_0^{\pi/2} d\theta = 1.$$

Hence

$$f_Z(z) = \frac{z}{\sigma^2}e^{-z^2/2\sigma^2}u(z),$$

which is the same result as obtained in Equation 3.3-33, obtained there by a different method.

Example 3.5-5

(*sum and difference again*) Finally, let us return to the problem considered in Example 3.4-1:

$$V \triangleq X + Y$$

$$W \triangleq X - Y.$$

The only root to

$$v - (x + y) = 0$$

$$w - (x - y) = 0$$

is

$$x = \frac{v + w}{2}$$

$$y = \frac{v - w}{2}$$

and $|\tilde{J}| = \frac{1}{2}$. Hence

$$f_{VW}(v,w) = \frac{1}{2} f_{XY}\left(\frac{v+w}{2}, \frac{v-w}{2}\right).$$

We verify in passing that

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} f_{VW}(v,w)dw \\ &= \int_{-\infty}^{\infty} \frac{1}{2} f_{XY}\left(\frac{v+w}{2}, \frac{v-w}{2}\right) dw, \qquad \text{with} \quad z \triangleq \frac{v+w}{2} \\ &= \int_{-\infty}^{\infty} f_{XY}(z, v-z)dz. \end{aligned}$$

This important result on the sum of two RVs was derived in Section 3.3 (Equation 3.3-12) by different means.

SUMMARY

The material in this chapter discussed *functions of random variables*, a subject of great significance in applied science and engineering and fundamental to the study of random processes. The basic problem dealt with computing the probability law of an *output* random variable Y produced by a system transformation g operating on an *input* random variable X (i.e., $Y = g(X)$). The problem was then extended to two input random variables X, Y being operated upon by system transformations g and h to produce two output random variables $V = g(X,Y)$ and $W = h(X,Y)$. Then the problem is to compute the joint pdf (PMF) of V, W from the joint pdf (PMF) of X, Y.

We showed how most problems involving functions of RVs could be computed in at least two ways:

1. the so-called indirect approach through the CDF; and
2. directly through the use of a "turn-the-crank" direct method.

A number of important problems involving transformations of random variables were worked out including computing the pdf (and PMF) of the sum of two random variables, a problem which has numerous applications in science and engineering where unwanted additive noise contaminates a desired signal or measurement. For example, the so-called "signal and additive noise problem" is a seminal issue in communications engineering.

Later, when we extend the analysis of the sum of two independent random variables to the sum of n independent random variables, we will begin to observe that the CDF of the sum starts to "look like" the CDF of a Normal random variable. This fundamental result, that is, convergence to the CDF of the Normal, is called the *Central Limit Theorem*, and is discussed in Chapter 4.

Finally we considered how to compute the pdf of two ordered random variables. We found we could do this using the powerful so-called direct method for computing distributions

of RVs fuctionally related to other RVs. Later, in Chapter 5 on random vectors, we will discuss transformations involving n ordered RVs. Ordered random variables appear in a branch of statistics called nonparametric statistics and often yield results that are independent of underlying distributions. In this sense, ordered random variables yield a certain level of robustness to expressions derived about them.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

3.1 Let X have CDF $F_X(x)$ and consider $Y = aX + b$, where $a < 0$. Show that if X is not a continuous RV, then Equation 3.2-3 should be modified to

$$F_Y(y) = 1 - F_X\left(\frac{y-b}{a}\right) + P\left[X = \frac{y-b}{a}\right]$$
$$= 1 - F_X\left(\frac{y-b}{a}\right) + P_X(\frac{y-b}{a}).$$

3.2 Let Y be a function of the RV X as follows:

$$Y \triangleq \begin{cases} X, & X \geq 0, \\ X^2, & X \leq 0. \end{cases}$$

Compute $f_Y(y)$ in terms of $f_X(x)$. Assume that X:$N(0,1)$.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

3.3 Let X be a random variable uniform over

(a) $(-\pi/2, \pi/2)$. Compute the pdf of $Y = \tan X$.
(b) (0, 1). Compute the pdf of $Y = e^X$.

3.4 Let Y be a function of the random variable X as follows:

$$Y \triangleq \begin{cases} X, & X \geq 0, \\ 2X^2, & X < 0. \end{cases}$$

Compute pdf $f_Y(y)$ in terms of pdf $f_X(x)$. Let $f_X(x)$ be given by

$$f_X(x) = \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{4}x^2},$$

that is, X:$N(0,2)$.

3.5 Let X have pdf

$$f_X(x) = \alpha e^{-ax}u(x).$$

Compute the pdf of (a) $Y = X^3$; (b) $Y = 3X + 2$.

3.6 Let X be a Laplacian random variable with pdf

$$f_X(x) = \frac{1}{2}e^{-|x|}, \qquad -\infty < x < +\infty.$$

Let $Y = g(X)$, where $g(\cdot)$ is the nonlinear function given as the saturable limiter

$$g(x) \triangleq \begin{cases} -1, & x < -1, \\ x, & -1 \leq x \leq +1, \\ +1, & x > +1. \end{cases}$$

Find the distribution function $F_Y(y)$.

***3.7** In medical imaging such as computer tomography, the relation between detector readings y and body absorptivity x follows a $y = e^x$ law. Let $X{:}N(\mu, \sigma^2)$; compute the pdf of Y. This distribution of Y is called *lognormal*. The lognormal random variable has been found quite useful for modeling failure rates of semiconductors, among many other uses.

***3.8** In the previous problem you found that if $X : N(\mu, \sigma^2)$, then $Y \triangleq \exp X$ has a lognormal density or pdf

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left[-\frac{(\ln y - \mu)^2}{2\sigma^2}\right] u(y).$$

(a) Sketch the lognormal density for a couple of values of μ and σ.

(b) What is the distribution function of the lognormal random variable Y? Express your answer in terms of our erf function. Hint: There are two possible approaches. You can use the method of substitution to integrate the above density, or you can find the distribution function of Y directly as a transformation of random variable problem.

***3.9** In homomorphic image processing, images are enhanced by applying nonlinear transformations to the image functions. Assume that the image function is modeled as RV X and the enhanced image Y is $Y = \ln X$. Note that X cannot assume negative values. Compute the pdf of Y if X has an exponential density $f_X(x) = \frac{1}{3}e^{-\frac{1}{3}x}u(x)$.

3.10 Assume that $X{:}N(0, 1)$ and let Y be defined by

$$Y = \begin{cases} \sqrt{X}, & X \geq 0, \\ 0, & X < 0. \end{cases}$$

Compute the pdf of Y.

3.11 (a) Let $X{:}N(0, 1)$ and let $Y \triangleq g(X)$, where the function g is shown in Figure P3.11. Use the indirect approach to compute $F_Y(y)$ and $f_Y(y)$ from $f_X(x)$. (b) Compute

$f_Y(y)$ from Equation 3.2-23. Why can't Equation 3.2-23 be used to compute $f_Y(y)$ at $y = 0, 1$?

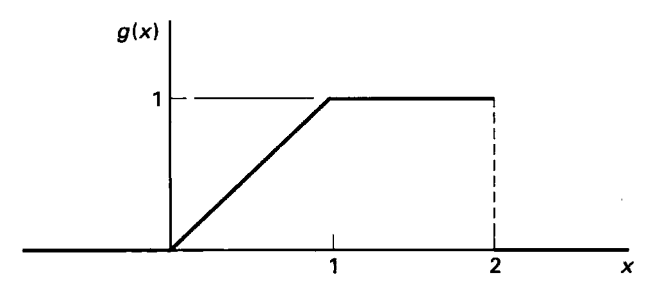


Figure P3.11

3.12 Let $X{:}U[0, 2]$. Compute the pdf of Y if $Y = g(X)$, where the function g is plotted in Figure P3.12.

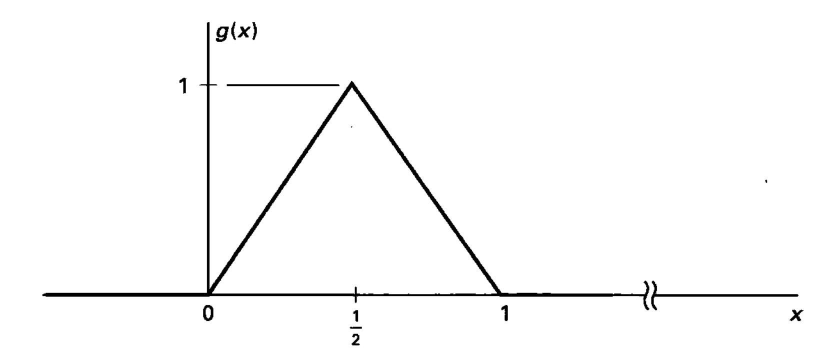


Figure P3.12

3.13 Let $X{:}U[0, 2]$, Compute the pdf of Y if $Y = g(X)$ with the function g as shown in Figure P3.13.

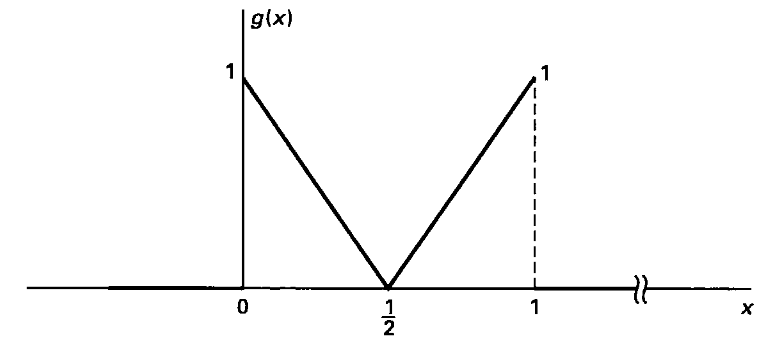


Figure P3.13

3.14 Let the RV $X{:}N(0,1)$, that is, X is Gaussian with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \qquad -\infty < x < +\infty.$$

Let $Y = g(X)$, where g is the nonlinear function given as

$$g(x) \triangleq \begin{cases} -1, & x < -1, \\ x, & -1 \le x \le 1, \\ 1, & x > 1. \end{cases}$$

It is called a *saturable limiter* function. (a) Sketch $g(x)$; (b) find $F_Y(y)$; (c) find and sketch $f_Y(y)$.

3.15 Let $X \sim N(\mu, \sigma^2)$ and let $Y = aX + b$. Show that $Y \sim N(a\mu + b, a^2\sigma^2)$ and find the values of a and b so that $Y \sim N(0,1)$.

3.16 Let $Y \triangleq \sec X$. Compute $f_Y(y)$ in terms of $f_X(x)$. What is $f_Y(y)$ when $f_X(x)$ is uniform in $(-\pi, \pi]$?

3.17 Consider two random variables X and Y with the joint pdf $f_{X,Y}(x,y)$. Determine the pdf of $Z = XY$. Repeat for the case when X and Y are independent uniform random variables over (0, 1).

3.18 Let X and Y be independent and identically distributed exponential RVs with

$$f_X(x) = f_Y(x) = \alpha e^{-\alpha x} u(x).$$

Compute the pdf of $Z \triangleq Y - X$.

3.19 Let random variables X and Y be described by the given joint pdf $f_{X,Y}(x,y)$. Define new random variables as

$$V \triangleq X + Y \quad \text{and} \quad W \triangleq 2X - Y.$$

(a) Find the joint pdf $f_{V,W}(v,w)$ in terms of the joint pdf $f_{X,Y}(x,y)$.
(b) Show, using the results of part (a) or in any other valid way, that under suitable conditions

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z - x) dx,$$

for $Z \triangleq X + Y$. What are the suitable conditions?

3.20 Repeat Example 3.2-11 for $f_X(x) = e^{-x}u(x)$.

3.21 Repeat Example 3.2-12 for $f_X(x) = e^{-x}u(x)$.

3.22 The objective is to generate numbers from the pdf shown in Figure P3.22. All that is available is a random number generator that generates numbers uniformly distributed in $(0,1)$. Explain what procedure you would use to meet the objective.

3.23 It is desired to generate zero-mean Gaussian numbers. All that is available is a random number generator that generates numbers uniformly distributed on $(0,1)$.

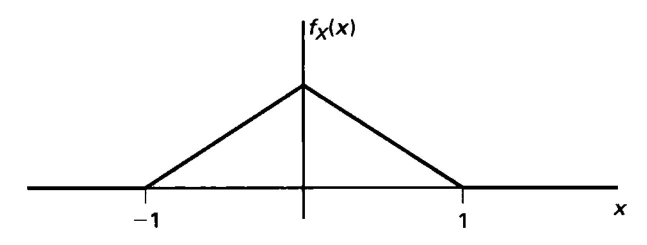


Figure P3.22

It has been suggested Gaussian numbers might be generated by adding 12 uniformly distributed numbers and subtracting 6 from the sum. Write a program in which you use the procedure to generate 10,000 numbers and plot a histogram of your result. A histogram is a bar graph that has bins along the x-axis and number of points in the bin along the y-axis. Choose 200 bins of width 0.1 to span the range from -10 to 10. In what region of the histogram does the data look most Gaussian? Where does it look least Gaussian? Give an explanation of why this approach works.

***3.24** Random number generators on computers often provide a basic uniform random variable $X\colon U[0,1]$. This problem explores how to get more general distributions by transformation of such an X.

(a) Consider the Laplacian density $f_Y(y) = \frac{c}{2}\exp(-c|y|)$, $-\infty < y < +\infty$, with parameter $c > 0$, that often arises in image processing problems. Find the corresponding Laplacian distribution function $F_Y(y)$ for $-\infty < y < +\infty$.

(b) Consider the transformation

$$z = g(x) = F_Y^{-1}(x),$$

using the distribution function you found in part (a). Note that F_Y^{-1} denotes an inverse function. Show that the resulting random variable $Z = g(X)$ will have the Laplacian distribution with parameter c if $X\colon U[0,1]$. Note also that this general result does not depend on the Laplacian distribution function other than that it has an inverse.

(c) What are the limitations of this transform approach? Specifically, will it work with mixed random variables? Will it work with distribution functions that have flat regions? Will it work with discrete random variables?

3.25 In Problem 3.18 compute the pdf of $|Z|$.

3.26 Let X and Y be independent, continuous RVs. Let $Z = \min(X, Y)$. Compute $F_Z(z)$ and $f_Z(z)$. Sketch the result if X and Y are distributed as $U(0,1)$. Repeat for the exponential density $f_X(x) = f_Y(x) = \alpha\exp[-\alpha x]\cdot u(x)$.

3.27 Let X and Y be two random variables with the joint pdf $f_{XY}(x,y)$ and joint CDF $F_{XY}(x,y)$. Let $Z = \max(X, Y)$.

(a) Find the CDF of Z.

(b) Find the pdf of Z if X and Y are independent.

Discuss if X and Y are independent and identical exponential variates with mean μ.

3.28 Let X and Y be two random variables with a joint pdf $f_{X,Y}(x, y)$.
Let $R = \sqrt{X^2 + Y^2}$, $\Theta = \tan^{-1}(Y/X)$. Find $f_{R\Theta}(r, \theta)$ in terms of $f_{XY}(x, y)$.

3.29 Let X, Y and Z be independent standard normal random variables.
Let $W = (X^2 + Y^2 + Z^2)^{1/2}$. Find the pdf of W.

3.30 Let $X_1, X_2, \ldots, X_n$ be n i.i.d. exponential random variables with $f_{X_i}(x) = e^{-x}u(x)$. Compute an explicit expression for the pdf of $Z_n = \max(X_1, X_2, \ldots, X_n)$. Sketch the pdf for $n = 3$.
Let $X_1, X_2, \ldots, X_n$ be n i.i.d. exponential random variables with $f_{X_i}(x) = e^{-x}u(x)$. Compute an explicit expression for the pdf of $Z_n = \min(X_1, X_2, \ldots, X_n)$. Sketch the pdf for $n = 3$.

3.31 Let X, Y be i.i.d. as $U(-1, 1)$. Compute and sketch the pdf of Z for the system shown in Figure P3.31. The square-root operation is valid only for positive numbers. Otherwise the output of the $\sqrt{\ }$ is zero.

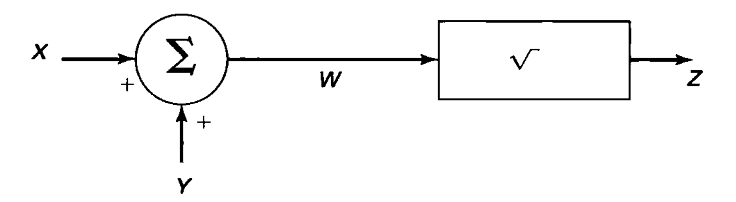


Figure P3.31 A square-root device.

3.32 The length of time, Z, an airplane can fly is given by $Z = \alpha X$, where X is the amount of fuel in its tank and $\alpha > 0$ is a constant of proportionality. Suppose a plane has two independent fuel tanks so that when one gets empty the other switches on automatically. Because of lax maintenance a plane takes off with neither of its fuel tanks checked. Let X_1 be the fuel in the first tank and X_2 the fuel in the second tank. Let X_1 and X_2 be modeled as uniform i.i.d. RVs with pdf $f_{X_1}(x) = f_{X_2}(x) = \frac{1}{b}[u(x) - u(x - b)]$. Compute the pdf of Z, the maximum flying time of the plane. If $b = 100$, say in liters, and $\alpha = 1$ hour/10 liters, what is the probability that the plane will fly *at least* five hours?

3.33 Let X and Y be two independent Poisson RVs with PMFs

$$P_X(k) = \frac{1}{k!}e^{-2}2^k u(k) \quad \text{and} \tag{3.5-6}$$

$$P_Y(k) = \frac{1}{k!}e^{-3}3^k u(k), \text{ respectively.} \tag{3.5-7}$$

Compute $P[Z \leq 4]$, where $Z \triangleq X + Y$. $\left[\textit{Hint: } \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j} = (a + b)^n.\right]$

3.34 Given two random variables X and Y that are independent and uniformly distributed as $U(0,1)$:

(a) Find the joint pdf $f_{U,V}$ of random variables U and V defined as:

$$U \triangleq \frac{1}{2}(X+Y) \text{ and}$$

$$V \triangleq \frac{1}{2}(X-Y).$$

(b) Sketch the support of $f_{U,V}$ in the (u,v) plane. Remember *support of a function* is the subset of its domain for which the function takes on nonzero values.

3.35 Let X and Y be independent random variables with pdf $f_X(x) = e^{-x}, x > 0$ and zero otherwise, and $f_Y(y) = e^{-y}, y > 0$ and zero otherwise. Compute (a) the pdf of $Z = \frac{X+Y}{2}$ (b) the pdf of $Z = X - Y$.

3.36 Compute the joint pdf $f_{ZW}(z,w)$ if

$$Z \triangleq X^2 + Y^2$$

$$W \triangleq X$$

when

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2} e^{-[(x^2+y^2)/2\sigma^2]}, -\infty < x < \infty, -\infty < y < \infty.$$

Then compute the $f_Z(z)$ from your results.

***3.37** Consider the transformation

$$Z = aX + bY$$

$$W = cX + dY.$$

Let

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} e^{-Q(x,y)},$$

where

$$Q(x,y) = \frac{1}{2\sigma^2(1-\rho^2)}[x^2 - 2\rho xy + y^2].$$

What combination of coefficients a, b, c, d will enable Z, W to be independent Gaussian RVs?

3.38 Let

$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\left(\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)\right].$$

Compute the joint pdf $f_{VW}(v,w)$ of

$$V = \tfrac{1}{2}(X^2 + Y^2) \tag{3.5-8}$$

$$W = \tfrac{1}{2}(X^2 - Y^2). \tag{3.5-9}$$

3.39 Derive Equation 3.4-4 by the direct method, that is, using Equations 3.4-11 or 3.4-12.

3.40 Consider the transformation

$$Z = X\cos\theta + Y\sin\theta \tag{3.5-10}$$

$$W = X\sin\theta - Y\cos\theta. \tag{3.5-11}$$

Compute the joint pdf $f_{ZW}(z, w)$ in terms of $f_{XY}(x, y)$ if

$$f_{XY}(x, y) = \frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}, -\infty < x < \infty, -\infty < y < \infty.$$

(It may be helpful to note that this transformation is a rotation by $+\theta$ followed by a negation on W.)

3.41 Compute the joint pdf of

$$Z \triangleq X^2 + Y^2$$

$$W \triangleq 2Y$$

when

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma^2}e^{-[(x^2+y^2)/2\sigma^2]}.$$

3.42 If X and Y are independent random variables which are uniformly distributed over (0, 1), find the joint pdf and hence the marginal pdf of $U = X + Y, V = X - Y$.

3.43 Consider the input–output view mentioned in Section 3.1. Let the underlying experiment be observations on an RV X, which is the input to a system that generates an output $Y = g(X)$.

(a) What is the range of Y?
(b) What are reasonable probability spaces for X and Y?
(c) What subset of R^1 consists of the event $\{Y \leq y\}$?
(d) What is the inverse image under Y of the event $(-\infty, y)$ if $Y = 2X + 3$?

3.44 In the diagram shown in Figure P3.44, it is attempted to deliver the signal X from points a to b. The two links L1 and L2 operate independently, with times-to-failure T_1, T_2, respectively, which are exponentially and identically distributed with rate λ (>0). Set $Y = 0$ if both links fail. Denote the output by Y and compute $F_Y(y, t)$, the CDF of Y at time t. Show for any fixed t that $F_Y(\infty, t) = 1$.

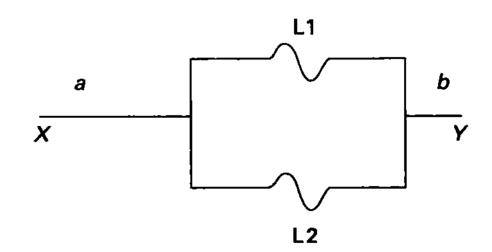


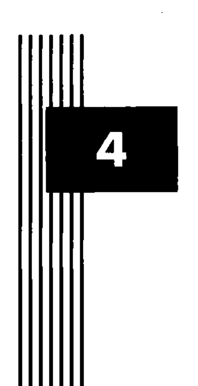
Figure P3.44 parallel links.

REFERENCES

3-1. W. F. Davenport, *Probability and Random Processes: An Introduction for Applied Scientists and Engineers.* New York: McGraw-Hill, 1970.
3-2. H. Stark, F. B. Tuteur, and J. B. Anderson, *Modern Electrical Communications.* 2nd edition, Upper Saddle River, N.J.: Prentice-Hall, 1988.
3-3. S. O. Rice, "Mathematical Analysis of Random Noise," *Bell System Technical Journal,* Vols. 23, 24, 1944, 1945.
3-4. J. Marsden and A. Weinstein, *Calculus.* Menlo Park, CA.: Benjamin/Cummings, 1980.
3-5. A. Papoulis, and S. U. Pillai *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 4th Ed, 2002.

ADDITIONAL READING

Cooper, G. R. and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis,* 3rd edition. New York: Holt, Rinehart and Winston, 1999.
Leon-Garcia, A., *Probability, Statistics, and Random Processes for Electrical Engineering,* 3rd edition. Reading, MA: Prentice Hall, 2008.
Helstrom, C. W., *Probability and Stochastic Processes for Engineers,* 2nd edition. New York: Macmillan, 1991.
Papoulis, A., *Probability & Statistics.* Englewood Cliffs, NJ: Prentice Hall, 1990.
Peebles, P. Z. Jr., *Probability, Random Variables, and Random Signal Principles,* 4th edition. New York: McGraw-Hill, 2001.
Scheaffer, R. L., *Introduction to Probability and Its Applications.* Belmont, CA: Duxbury, 1990.
Strook, D. W., Probability Theory, an Analytic View, 2nd edition, Cambridge University Press, Cambridge, England 2010.
Viniotis, Y., *Probability and Random Processes for Electrical Engineers.* New York: McGraw-Hill, 1998.
Yates, R. D. and D. J. Goodman, *Probability and Stochastic Processes,* 2nd edition, New York: Wiley, 2004.
Ziemer, R. E., *Elements of Engineering Probability & Statistics.* Upper Saddle River, NJ: Prentice Hall, 1997.



4 Expectation and Moments

4.1 EXPECTED VALUE OF A RANDOM VARIABLE

It is often desirable to summarize certain properties of an RV and its probability law by a few numbers. Such numbers are furnished to us by the various averages, or *expectations* of an RV; the term *moments* is often used to describe a broad class of averages, and we shall use it later.

We are all familiar with the notion of the average of a set of numbers, for example, the average class grade for an exam, the average height and weight of children at age five, the average lifetime of men versus women, and the like. Basically, we compute the average of a set of numbers $x_1, x_2, \ldots, x_N$ as follows:

$$\mu_s = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{4.1-1}$$

where the subscript s is a reminder that μ_s is the average of a set.

The average μ_s of a set of numbers $x_1, x_2, \ldots, x_N$ can be viewed as the "center of gravity" of the set. More precisely the average is the number that is simultaneously closest to all the numbers in the set in the sense that the sum of the distances from it to all the points in the set is smallest. To demonstrate this we need only ask what number z minimizes the summed distance D or summed distance-square D^2 to all the points. Thus with

$$D^2 \triangleq \sum_{i=1}^{N}(z - x_i)^2,$$

the minimum occurs when $dD^2/dz = 0$ or

$$\frac{dD^2}{dz} = 2Nz - 2\sum_{i=1}^{N} x_i = 0,$$

which implies that

$$z = \mu_s = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Note that each number in Equation 4.1-1 is given the same weight (i.e., each x_i is multiplied by the same factor $1/N$). If, for some reason we wish to give some numbers more weight than others when computing the average, we then obtain a *weighted* average. However, we won't pursue the idea of a weighted average any further in this chapter.

Although the average as given in Equation 4.1-1 gives us the "most likely" value or the "center of gravity" of the set, it does not tell us how much the numbers spread or deviate from the average. For example, the sets of numbers $S_1 = \{0.9, 0.98, 0.95, 1.1, 1.02, 1.05\}$ and $S_2 = \{0.2, -3, 1.8, 2, 4, 1\}$ have the same average but the spread of the numbers in S_2 is much greater than that of S_1. An average that summarizes this spread is the *standard deviation of the set*, σ_s, computed from

$$\sigma_s = \left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_s)^2\right]^{1/2}. \tag{4.1-2}$$

Equations 4.1-1 and 4.1-2, important as they are, fall far short of disclosing the usefulness of averages. To exploit the full range of applications of averages, we must develop a calculus of averages from probability theory.

Consider a probability space $(\Omega, \mathscr{F}, P)$ associated with an experiment $\mathscr{H}$ and a discrete RV X. Associated with each outcome ζ_i of $\mathscr{H}$, there is a value $X(\zeta_i) \triangleq x_i$, which the RV X takes on. Let $x_1, x_2, \ldots, x_M$ be the M distinct values that X can take. Now assume that $\mathscr{H}$ is repeated N times and let $x^{(k)}$ be the observed outcome at the kth trial. Note that $x^{(k)}$ must assume one of the numbers $x_1, \ldots, x_M$. Suppose that in the N trials x_1 occurs n_1 times, x_2 occurs n_2 times, and so forth. Then for N large, we can estimate the average value μ_X of X from the formula

$$\mu_X \simeq \frac{1}{N}\sum_{k=1}^{N} x^{(k)} \tag{4.1-3}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{M} n_i x_i$$

$$\simeq \sum_{i=1}^{M} x_i \left(\frac{n_i}{N}\right) \quad . \tag{4.1-4}$$

$$\simeq \sum_{i=1}^{M} x_i P[X = x_i]. \tag{4.1-5}$$

Example 4.1-1

(*loaded dice*) We observe 17 tosses of a loaded die. Here $N = 17$, $M = 6$ (the six faces of the die.) The observations are $\{1,3,3,1,2,1,3,2,1,1,2,4,1,1,5,3,6\}$. Let $P[i]$ denote the probability of observing the face with the number i on it. Then from the observational data we get

$$P[1] \simeq 7/17; P[2] \simeq 3/17; P[3] \simeq 4/17; P[4] \simeq 1/17; P[5] \simeq 1/17; P[6] = 1/17.$$

These estimates of the "true" probabilities are quite unreliable, however. To get more reliable date we would have to greatly increase the number of tosses. We might ask what are the "true " probabilities anyway. One answer might be that the Laws of Nature have imbued the die with an inherent set of probabilities that must be determined by experimentation. Another view is that the true probabilities are the ratios $P[i] = n_i/N$ you get when N becomes arbitrarily large. However what is meant by *arbitrarily large*? For any finite values of N the estimated probabilities will always change as we increase N. These conundrums are mostly resolved by *statistics* discussed in some detail Chapters 6 and 7.

Equation 4.1-5, which follows from the frequency definition of probability, leads us to our first definition.

Definition 4.1-1 The *expected* or *average* value of a discrete RV X taking on values x_i with PMF $P_X(x_i)$ is defined by

$$E[X] \triangleq \sum_i x_i P_X(x_i). \quad \blacksquare \tag{4.1-6}$$

As given, the expectation is computed in the probability space generated by the RV. We can also compute the expectation by summing over all points of the discrete sample space, that is, $E[X] = \sum_\Omega X(\zeta_i)P[\{\zeta_i\}]$, where the ζ_i are the discrete outcome points in the sample space Ω.

A definition that applies to both continuous and discrete RVs is the following:

Definition 4.1-2 The expected value or mean, if it exists,† of a real RV X with pdf $f_X(x)$ is defined by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx. \quad \blacksquare \tag{4.1-7}$$

Here, as well as in Definition 4.1-1, the expectation can be computed in the original probability space. If the sample description space is not discrete but continuous, for example, an

†The expected value will exist if the integral is absolutely convergent, that is, if $\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$.

uncountable infinite set of outcomes such as the real line. Then $E[X] = \int_\Omega X(\zeta)P[\{d\zeta\}]$, where $P[\{d\zeta\}]$ is the probability of the infinitesimal event $\{\zeta < \zeta' \leq \zeta + d\zeta\}$.

The symbols $E[X]$, $\overline{X}$, μ_X, or simply μ are often used interchangeably for the expected value of X. Consider now a function of an RV, say, $Y = g(X)$. The expected value of Y is, from Equation 4.1-7,

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy. \tag{4.1-8}$$

However, Equation 4.1-8 requires computing $f_Y(y)$ from $f_X(x)$. If all we want is $E[Y]$, is there a way to compute it without first computing $f_Y(y)$? The answer is given by Theorem 4.1-1 which follows.

Theorem 4.1-1 The expected value of $Y = g(X)$ can be computed from

$$E[Y] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx, \tag{4.1-9}$$

where g is a measurable (Borel) function.† Equation 4.1-9 is an important result in the theory of probability. A rigorous proof of Equation 4.1-9 requires some knowledge of Lebesgue integration; we offer instead an informal argument below to argue that Equation 4.1-9 is valid.‡

On the Validity of Equation 4.1-8

Recall from Section 3.2 that if $Y = g(X)$ then for any y_j (Figure 4.1-1)

$$\{y_j < Y \leq y_j + \Delta y_j\} = \bigcup_{k=1}^{r_j} \{x_j^{(k)} < X \leq x_j^{(k)} + \Delta x_j^{(k)}\}, \tag{4.1-10}$$

where r_j is the number of real roots of the equation $y_j - g(x) = 0$, that is,

$$y_j = g(x_j^{(1)}) = \ldots = g(x_j^{(r_j)}). \tag{4.1-11}$$

The equal sign in Equation 4.1-10 means that the underlying event is the same for both mappings X and Y. Hence the probabilities of the events on either side of the equal sign are equal. The events on the right side of Equation 4.1-10 are disjoint and therefore the probability of the union is the sum of the probabilities of the individual events. Now partition

†See definition of a measurable function in Section 3.1.

‡See Feller [4-1, p.5] or Davenport [4-2, p.223]

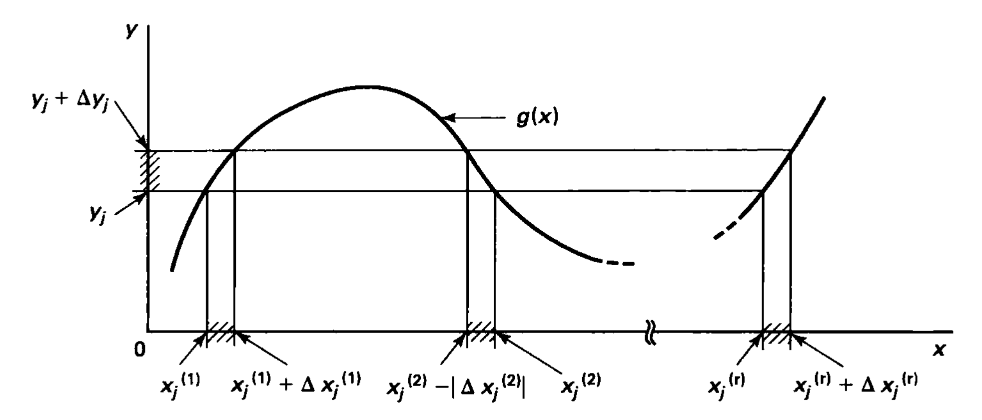


Figure 4.1-1 Equivalence between the events given in Equation 4.1-10.

the y-axis into many fine subintervals $y_1, y_2, \ldots, y_j, \ldots$. Then, approximating Equation 4.1-8 with a Riemann† sum and using Equation 2.4-6, we can write‡

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} y f_Y(y) dy \\ &\simeq \sum_{j=1}^{m} y_j P[y_j < Y \leq y_j + \Delta y_j] \\ &= \sum_{j=1}^{m} \sum_{k=1}^{r_j} g(x_j^{(k)}) P[x_j^{(k)} < X \leq x_j^{(k)} + \Delta x_j^{(k)}]. \end{aligned} \tag{4.1-12}$$

The last line of Equation 4.1-12 is obtained with the help of Equations 4.1-10 and 4.1-11. But the points $x_j^{(k)}$ are distinct, so that the cumbersome double indices j and k can be replaced with a single subscript index, say, i, The Equation 4.1-12 becomes

$$E[Y] \simeq \sum_{i=1}^{n} g(x_i) P[x_i < X \leq x_i + \Delta x_i],$$

and as Δy, $\Delta x \to 0$ we obtain the exact result that

$$E[Y] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{4.1-13}$$

Equation 4.1-13 follows from the Riemann sum approximation and Equation 2.4-6; the x_i have been ordered in increasing order $x_1 < x_2 < x_3 < \ldots$.

†Bernhard Riemann (1826–1866). German mathematician who made numerous contributions to the theory of integration.

‡The argument follows that of Papoulis [4-3, p.141]

In the special case where X is a discrete RV,

$$E[Y] = \sum_i g(x_i)P_X(x_i). \tag{4.1-14}$$

This result follows immediately from Equation 4.1-13, since the pdf of a discrete RV involves delta functions that have the property given in Equation B.3-1 in Appendix B. ■

Example 4.1-2

(*expected value of Gaussian*) Let $X: N(\mu, \sigma^2)$, read "X is distributed as Normal with parameters μ and σ^2." The expected value or mean of X is

$$E[X] = \int_{-\infty}^{\infty} x \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right) \right) dx.$$

Let $z \triangleq (x - \mu)/\sigma$. Then

$$E[X] = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{1}{2}z^2} dz + \mu \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \right).$$

The first term is zero because the integrand is odd, and the second term is μ because the term in parentheses is $P[Z \leq \infty]$, which is the certain event for $Z: N(0,1)$. Hence

$$E[X] = \mu \qquad \text{for } X: N(\mu, \sigma^2).$$

Thus, the parameter μ in $N(\mu, \sigma^2)$ is indeed the expected or mean value of X as claimed in Section 2.4.

Example 4.1-3

(*expected value of Bernoulli RV*) Assume that the RV B is Bernoulli distributed taking on value 1 with probability p and 0 with probability $q = 1 - p$. Then the PMF is given as

$$P_B(k) = \begin{cases} p, & \text{when } k = 1, \\ q, & \text{when } k = 0, \\ 0, & \text{else.} \end{cases}$$

The expected value is then given as

$$\begin{aligned} E[B] &= \sum_{b=-\infty}^{+\infty} k P_B(k) \\ &= 1p + 0 \\ &= p. \end{aligned}$$

Example 4.1-4

(*expected value of binomial RV*) Assume that the RV K is binomial distributed with PMF $P_K(k) = b(k; n, p)$. Then we calculate the expected value as

$$
\begin{aligned}
E[K] &= \sum_{k=-\infty}^{+\infty} kP_K(k) \\
&= \sum_{k=0}^{n} kb(k; n, p) \\
&= \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= \sum_{k'=0}^{n-1} \frac{n!}{k'!(n-k'-1)!} p^{k'+1} (1-p)^{n-k'-1} \qquad \text{with} \quad k' \triangleq k-1, \\
&= np\left(\sum_{k'=0}^{n-1} \frac{(n-1)!}{k'!(n-1-k')!} p^{k'} (1-p)^{n-1-k'}\right) \\
&= np, \quad \text{since the sum in the round brackets is } \sum_{k=0}^{n-1} b(k; n-1, p) = 1.
\end{aligned}
$$

Example 4.1-5

(*more on multiple lottery tickets*) We continue Example 1.9-6 of Chapter 1 on whether it is better to buy 50 tickets from a single lottery or 1 ticket each from 50 successive lotteries, all independent and with the same fair odds. Here we are interested in the mean or expected return in each case. Again each lottery has 100 tickets at \$1 each and the fair payoff is \$100 to the winner. For the single lottery, we remember the odds of winning are 50 percent, so the expected payoff is \$50. For the 50 plays in separate lotteries, we recall that the number of wins K is binomial distributed as $b(k; 50, 0.01)$, so the mean value $E[K] = np = 50 \times 0.01 = 0.5$. Since the payoff would be \$100K, the average payoff would be \$50, same as in the single lottery.

Example 4.1-6

(*expected value of Poisson*) Let K be a Poisson RV with parameter $a > 0$. Then

$$E[K] = \sum_{k=0}^{\infty} k\frac{e^{-a}}{k!}a^k$$

$$= a\sum_{k=0}^{\infty}\frac{e^{-a}}{(k-1)!}a^{k-1}$$

$$= a\sum_{k=1}^{\infty}\frac{e^{-a}}{(k-1)!}a^{k-1}$$

$$= a\sum_{i=0}^{\infty}\frac{e^{-a}}{i!}a^{i}$$

$$= a. \tag{4.1-15}$$

Thus, the expected value of Poisson RV is the parameter a.

Linearity of expectation. When we regard mathematical expectation E as a operator, it is relatively easy to see it is a linear operator. For any X, consider

$$E\left[\sum_{i=1}^{N} g_i(X)\right] = \int_{-\infty}^{+\infty}\left(\sum_{i=1}^{N} g_i(x)\right) f_X(x)dx \tag{4.1-16}$$

$$= \sum_{i=1}^{N}\int_{-\infty}^{+\infty} g_i(x)f_X(x)dx \tag{4.1-17}$$

$$= \sum_{i=1}^{N} E[g_i(X)] \tag{4.1-18}$$

provided that these exist. The expectation operator E is also linear for the sum of two RVs:

$$E[X+Y] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{+\infty} x\left(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy\right)dx + \int_{-\infty}^{+\infty} y\left(\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx\right)dy$$

$$= \int_{-\infty}^{+\infty} xf_X(x)dx + \int_{-\infty}^{+\infty} yf_Y(x)dy$$

$$= E[X] + E[Y].$$

The reader will notice that this result can readily be extended to the sum of N RVs $X_1, X_2, \ldots, X_N$. Thus,

$$E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i]. \tag{4.1-19}$$

Note that *independence is not required.* We can summarize both linearity results by saying that the mathematical expectation operator E distributes over a sum of RVs.

Example 4.1-7

(*variance of Gaussian*) Let $X : N(\mu, \sigma^2)$ and consider the zero-mean RV $X - \mu$ with variance $E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - \mu^2 = \text{Var}[X]$ by the linearity of expectation E. We can write

$$\begin{aligned} E[(X-\mu)^2] &= \int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z^2 e^{-\frac{z^2}{2}} dz \quad \text{with substitution } z \triangleq (x-\mu)/\sigma. \end{aligned}$$

Next we integrate by parts with $u = z$ and $dv = ze^{-\frac{z^2}{2}} dz$, yielding $du = dz$ and $v = -e^{-\frac{z^2}{2}}$, so that, the above integral becomes

$$\begin{aligned} \int_{-\infty}^{+\infty} z^2 e^{-\frac{z^2}{2}} dz &= \left(-ze^{-\frac{z^2}{2}}\right) \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \\ &= -0 + 0 + \sqrt{2\pi}, \end{aligned}$$

where the last term is due to the fact that the standard Normal $N(0,1)$ density integrates to 1. Thus we have $E[(X-\mu)^2] = \frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi} = \sigma^2$, and thus the parameter σ^2 in the Gaussian density is shown to be the variance of the RV $X - \mu$, which is the same as the variance of the RV X.

We have now established that the parameters introduced in Chapter 2, upon definition of the Gaussian density, are actually the mean and variance of this distribution. In practice these basic parameters are often estimated by making many independent observations on X and using Equation 4.1-1 to estimate the mean and Equation 4.1-2 to estimate σ.

Example 4.1-8

(*mean of Cauchy*) The Cauchy pdf with parameters $\alpha(-\infty < \alpha < \infty)$ and $\beta(\beta > 0)$ is given by

$$f_X(x) = \frac{1}{\pi\beta\left(1 + \left(\frac{x-\alpha}{\beta}\right)^2\right)}, \qquad -\infty < x < \infty. \tag{4.1-20}$$

Let X be Cauchy with $\beta = 1$, $\alpha = 0$. Then

$$E[X] = \int_{-\infty}^{\infty} x \left(\frac{1}{\pi(x^2+1)}\right) dx$$

is an improper integral and doesn't converge in the ordinary sense. However, if we evaluate the integral in the Cauchy principal value sense, that is,

$$E[X] = \lim_{x_0 \to \infty} \left[\int_{-x_0}^{x_0} x \left(\frac{1}{\pi(x^2+1)} \right) dx \right], \tag{4.1-21}$$

then $E[X] = 0$. Note, however, that with $Y \triangleq X^2$, $E[Y]$ doesn't exist in any sense because

$$E[Y] = \int_{-\infty}^{\infty} x^2 \left[\frac{1}{\pi(x^2+1)} \right] dx = \infty \tag{4.1-22}$$

and thus fails to converge in any sense. Thus, the variance of a Cauchy RV is infinite.

Expected value of a function of RVs. For a function of two RVs, that is, $Z = g(X, Y)$, the expected value of Z can be computed from

$$E[Z] = \int_{-\infty}^{\infty} z f_Z(z) dz$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dx\, dy. \tag{4.1-23}$$

To prove that Equation 4.1-23 can be used to compute $E[Z]$ requires an argument similar to the one we used in establishing Equation 4.1-9. Indeed one would start with an equation very similar to Equation 4.1-10, for example,

$$\{z_j < Z \leq z_j + \Delta z\} = \bigcup_{k=1}^{N_j} \{(X, Y) \in D_k\},$$

where the D_k are very small disjoint regions containing the points $(x_k^{(j)}, y_k^{(j)})$ such that $g(x_k^{(j)}, y_k^{(j)}) = z_j$. Taking probabilities of both sides and recalling that the D_k are disjoint, yields

$$f_Z(z_j)\Delta z_j \simeq \sum_{k=1}^{N_j} f(x_k^{(j)}, y_k^{(j)}) \Delta a_k^{(j)},$$

where $\Delta a_k^{(j)}$ is an infinitesimal area.

Now multiply both sides by z_j and recall that $z_j = g(x_k^{(j)}, y_k^{(j)})$. Then

$$z_j f_Z(z_j)\Delta z_j \simeq \sum_{k=1}^{N_i} g(x_k^{(j)}, y_k^{(j)}) f_{XY}(x_k^{(j)}, y_k^{(j)}) \Delta a_k^{(j)}$$

and, as $j \to \infty$, $\Delta z_j \to 0$, $\Delta a_k^{(j)} \to da = dx\, dy$,

$$\int_{-\infty}^{\infty} z f_Z(z) dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y)\, dx\, dy.$$

An alternative proof is of interest[†]. As before let $Z = g(X,Y)$ and write

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} z f_Z(z) dz \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z f_{Z|Y}(z|y) f_Y(y) dy\, dz. \end{aligned}$$

The second line follows from the definition of a marginal pdf. Now recall that if $Z = g(X)$ then

$$\int_{-\infty}^{\infty} z f_Z(z) dz = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

We can use this result in the present problem as follows. If we hold Y fixed at $Y = y$, then $g(X, y)$ depends only on X, and the conditional expectation of z with $Y = y$ is

$$\int_{-\infty}^{\infty} z f_{Z|Y}(z|y) dz = \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y)\, dx.$$

Using this result in the above yields

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} z f_Z(z) dz \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} z f_{Z|Y}(z|y) dz \right) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x|y) f_Y(y)\, dx\, dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\, dx\, dy. \end{aligned}$$

Example 4.1-9

(*mean of product of independent RVs*) Let $g(x,y) = xy$. Compute $E[Z]$ if $Z = g(X,Y)$ with X and Y independent and Normal with pdf

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2} \left((x - \mu_a)^2 + (y - \mu_b)^2 \right) \right].$$

[†] Carl W. Helstrom, *Probability and Stochastic Processes for Engineers*, 2nd edition. New York, Macmillan, 1991.

Solution Direct substitution into Equation 4.1-23 and recognizing that the resulting double integral factors into the product of two single integrals enables us to write

$$\begin{aligned} E[Z] &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x\exp\left[-\frac{1}{2\sigma^2}(x-\mu_a)^2\right]dx \\ &\times \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} y\exp\left[-\frac{1}{2\sigma^2}(y-\mu_b)^2\right]dy \\ &= \mu_a\mu_b. \end{aligned}$$

Equation 4.1-23 can be used to compute $E[X]$ or $E[Y]$. Thus with $Z = g(X,Y) = X$, we obtain

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f_{XY}(x,y)dy\right]x\,dx. \end{aligned} \quad (4.1\text{-}24)$$

By Equation 2.6-47, the integral in brackets is the *marginal* pdf $f_X(x)$. Hence Equation 4.1-23 is completely consistent with the definition

$$E[X] \triangleq \int_{-\infty}^{\infty} x f_X(x)\,dx.$$

With the help of marginal densities we can conclude that

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{XY}(x,y)dxdy \\ &= \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f_{XY}(x,y)dy\right)dx + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f_{XY}(x,y)dx\right)dy \\ &= E[X]+E[Y]. \end{aligned} \quad (4.1\text{-}25)$$

Equation 4.1-24 can be extended to N random variables $X_1, X_2, \ldots, X_N$. Thus

$$E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} E[X_i] \quad (4.1\text{-}26)$$

Note that independence is not required.

Example 4.1-10

(*independent Normal RVs*) Let X, Y be jointly normal, independent RVs with pdf

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right).$$

It is clear that X and Y are independent since $f_{XY}(x, y) = f_X(x)f_Y(y)$. The marginal pdf's are obtained using Equations 2.6-44 and 2.6-47:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]$$

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

Thus Equation 4.1-24 yields

$$E[X+Y] = \mu_1 + \mu_2.$$

Example 4.1-11

(*Chi-square law*) In a number of problems in engineering and science, signals add incoherently, meaning that the power in the sum of the signals is merely the sum of the powers. This occurs, for example, in optics when a surface is illuminated by light sources of different wavelengths. Then the power measured on the surface is just the sum of the powers contributed by each of the sources. In electric circuits, when the sources are sinusoidal at different frequencies, the power dissipated in any resistor is the sum of the powers contributed by each of the sources. Suppose the individual source signals, at a given instant of time, are modeled as identically distributed Normal RVs. In particular let $X_1, X_2, \ldots, X_n$ represent the n independent signals produced by the n sources with X_i: $N(0,1)$ for $i = 1, 2, \ldots, n$ and let $Y_i = X_i^2$. We know from Example 3.2-2 in Chapter 3 that the pdf of Y_i is given by

$$f_{Y_i}(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} u(y).$$

Consider now the sums $Z_2 = Y_1 + Y_2$, $Z_3 = Y_1 + Y_2 + Y_3, \ldots, Z_n = \Sigma_{i=1}^n Y_i$. The pdf of Z_2 is easily computed by convolution as

$$\begin{aligned} f_{Z_2}(z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi x}} e^{-x/2} u(x) \times \frac{1}{\sqrt{2\pi(z-x)}} e^{-\frac{1}{2}(z-x)} u(z-x)\, dx \\ &= \frac{1}{\pi} e^{-z/2} \int_0^{\sqrt{z}} \frac{dy}{\sqrt{z-y^2}} u(z) \\ &= \frac{1}{2} e^{-z/2} u(z) \text{ (exponential pdf).} \end{aligned}$$

To get from line 1 to line 2 we let $x = y^2$. To get from line 2 to line 3, we used that the integral is an elementary trigonometric function integral in disguise. To get the pdf of Z_3 we convolve the pdf of Z_2 with that of Y_3. The result is

$$\begin{aligned} f_{Z_3}(z) &= \frac{1}{2} \int_{-\infty}^{\infty} e^{-x/2} u(x) \times \frac{1}{\sqrt{2\pi(z-x)}} e^{-\frac{1}{2}(z-x)} u(z-x)\, dx \\ &= \frac{1}{\sqrt{2\pi}} z^{\frac{1}{2}} e^{-\frac{1}{2}z} u(z). \end{aligned}$$

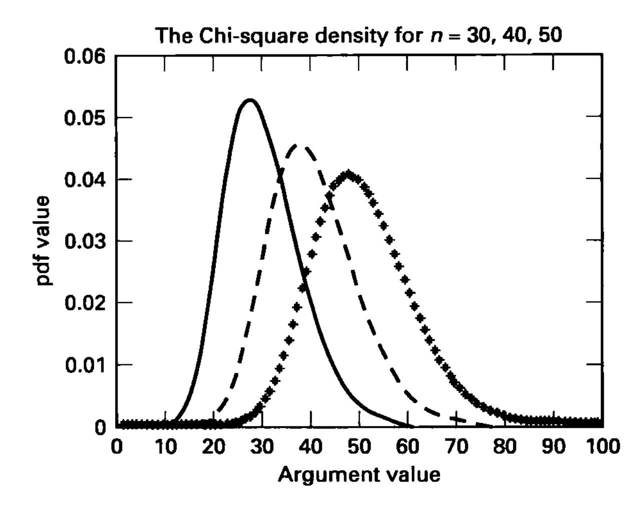


Figure 4.1-2 The Chi-square pdf for three large values for the parameter n: $n = 30$ (solid); $n = 40$ (dashed); $n = 50$ (stars). For large values of n, the Chi-square pdf can be approximated by a normal $N(n, 2n)$ for computing probabilities not too far from the mean. For example, for $n = 30$, $P[\mu - \sigma < X < \mu + \sigma] = 0.6827$ assuming X: $N(30, 60)$. The value computed, using single-precision arithmetic, using the Chi-square pdf, yields 0.6892.

We leave the intermediate steps which involve only elementary transformations to the reader. Proceeding in this way, or using mathematical induction, we find that

$$f_{Z_n}(z) = \frac{1}{2^{n/2}\,\Gamma(n/2)} z^{\frac{n-2}{2}} e^{-z/2} u(z).$$

This pdf was introduced in Chapter 2 as the Chi-square pdf. More precisely, it is known as the *Chi-square distribution with n degrees-of-freedom*. For $n > 2$, the pdf has value zero at $z = 0$, reaches a peak, and then exhibits monotonically decreasing tails. For large values of n, it resembles a Gaussian pdf with mean in the vicinity of n. However, the Chi-square can never be truly Gaussian because the Chi-square RV never takes on negative values. The character of the Chi-square pdf is shown in Figure 4.1-2 for different values of large n.

The mean and variance of the Chi-square RV are readily computed from the definition $Z_n \triangleq \Sigma_{i=1}^n X_i^2$. Thus $E[Z_n] = E[\Sigma_{i=1}^n X_i^2] = \Sigma_{i=1}^n E[X_i^2] = n$. Also $\text{Var}(Z_n) = E[(Z_n - n)^2]$. After simplifying, we obtain $\text{Var}(Z_n) = E[Z_n^2] - n^2$. We leave it to the reader to show that $E[Z_n^2] = 2n + n^2$ and, hence, that $\text{Var}(Z_n) = 2n$.

Example 4.1-12

At the famous University of Politicalcorrectness (U of P), the administration requires that each professor be equipped with an electronic Rolodex which contains the names of every student in the class. When the professor wishes to call on a student, she merely hits the "call" button on the Rolodex, and a student's name is selected randomly by an electronic circuit inside the Rolodex. By using this device the professor becomes immune to charges

of bias in the selection of students she calls on to answer her questions. Find an expression for the *average number* of "calls" r required so each student is called upon at least once.

Solution The use of the electronic Rolodex implies that some students may not be called at all during the entire semester and other students may be called twice or three times in a row. It will depend on how big the class is. Nevertheless the average is well defined because extremely long bad runs, that is, where one or more students are not called on, are very rare. The careful reader may have observed that this is an occupancy problem if we associate "calls" with *balls* and students with *cells*. Let $R \in \{n, n+1, n+2, \ldots\}$ denote number of balls needed to fill all the n cells *for the first time*. The only way that this can happen is that the first $R-1$ balls fill all but one of the n cells (event E_1) and the Rth ball fills the remaining empty cell (event E_2). Translated to the class situation, this means that after $R-1$ calls, all but one student will have been called (event E_1) and this student will be called on the Rth call (event E_2). Thus $P[R=r,n] \triangleq P_R(r,n) = P[E_1E_2] = P[E_1]P[E_2]$ since E_1 and E_2 are independent. Now $P[E_2] = 1/n$ since it is merely the probability that a given ball goes into a selected cell, and $P[E_1]$ is $P_1(r-1,n)$ of Equation 1.8-13, that is

$$P_1(r-1,n) = \binom{n}{1}\sum_{i=0}^{n-1}\binom{n}{i}(-1)^i\left(1-\frac{i+1}{n}\right)^{r-1}, \quad r \geq n$$
$$= 0, \text{else.}$$

Thus $P_R(r,n)$ is given by

$$P_R(r,n) = \sum_{i=0}^{n-1}\binom{n}{i}(-1)^i\left(1-\frac{i+1}{n}\right)^{r-1}, \quad r \geq n \qquad (4.1\text{-}27)$$
$$= 0, \text{else.}$$

The probability $P_K(k,n)$ that all n cells (students) have been filled (called) after distributing k balls (called k students) is, from Equation 1.8-9

$$P_K(k,n) = \sum_{i=0}^{n}\binom{n}{i}(-1)^i\left(1-\frac{i}{n}\right)^{k}, \quad k \geq n \qquad (4.1\text{-}28)$$
$$= 0, \text{else.}$$

Finally, the expected value of the RV R is given by

$$E[R] = \sum_{r=n}^{\infty} r\left(\sum_{i=0}^{n-1}\binom{n}{i}(-1)^i\left(1-\frac{i+1}{n}\right)^{r-1}\right) \qquad (4.1\text{-}29)$$

Example 4.1-13

Write a MATLAB program for computing the probability that all the students in Example 4.1-12 are called upon at least once in r calls from the electronic Rolodex. Assume there are 20 students in the class.

Solution The appropriate equation to be coded is Equation 4.1-28. The result is shown in Figure 4.1-3.

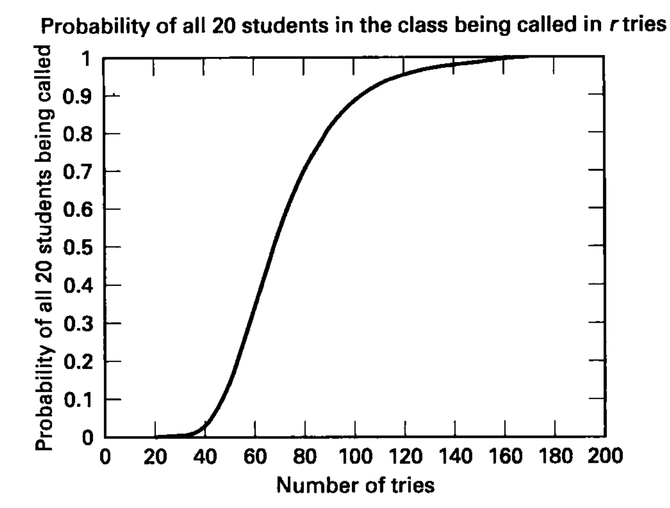


Figure 4.1-3 MATLAB result for Example 4.1-13.

```
function [tries,prob]=occupancy(balls,cells)
tries=1:balls; % identifies a vector ''tries''
prob=zeros(1,balls); % identifies a vector ''prob''
a=zeros(1,cells); % identifies a vector ''a''
d=zeros(1,cells); % identifies a vector ''d''

term=zeros(1,cells); % identifies a vector ''term''
% next follows the realization of Equation (4.1-27)
for m=1:balls

    for k=1:cells
        a(k)=(-1)^k)*prod(1:cells)/(prod(1:k)*prod(1:cells-k));
        d(k)=(1-(k/cells))^m;
        term(k)=a(k)*d(k);
    end
prob(m)=1+sum(term);
end
  plot(tries,prob)
  title(['Probability of all ' num2str(cells) ' students in the class
being called in r tries'])
xlabel('number of tries')
ylabel(['Probability of all ' num2str(cells) ' students being called
'])
```

Example 4.1-14

Write a MATLAB program for computing the average number of calls required for each student to be called at least once. Assume a maximum of 50 students and make sure the number of calls is large ($n \geq 400$).

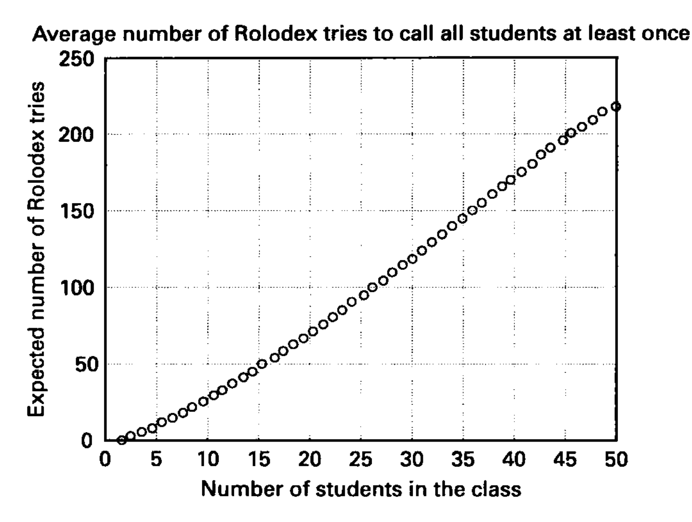


Figure 4.1-4 MATLAB result for Example 4.1-14.

Solution The appropriate equation to be coded is Equation 4.1-29. The result is shown in Figure 4.1-4.

```
function [cellvec,avevec]=avertries(ballimit,cellimit);
cellvec = 1:cellimit;
termvec = zeros(1,ballimit);
avevec = zeros(1,cellimit);
brterm=zeros(1,ballimit);
srterm=zeros(1,ballimit);
for n=1:cellimit;
   a = zeros(1,n);
   d = zeros(1,n);
   termvec = zeros(1,n);
   for r=1:ballimit
      for i=1:n-1
         a(i) = ((-1)^i)*prod(1:n-1)/(prod(1:i)*prod(1:n-1-i));
         d(i) = (1-((i-1)/n))^(r-1);
         termvec(i) = a(i)*d(i);
      end
      brterm(r)=r*sum(termvec);
      lrterm(r)=r*((1-(1/n)))^(r-1);
   end
   avevec(n)=sum(brterm)+sum(lrterm);
end
plot(cellvec,avec,'o')
title('Average number of Rolodex tries to call all students at least
once')
xlabel('number of students in the class')
```

```
ylabel('Expected number of Rolodex tries to reach all students at least
once')
grid
```

Example 4.1-15

(*geometric distribution*) The RV X is said to have a *geometric distribution* if its probability mass function is given by

$$P_X(n) = (1-a)a^n u(n),$$

where u is the unit-step function† and $0 < a < 1$. Clearly $\Sigma_{n=0}^{\infty} P_X(n) = 1$, a result easily obtained from $\Sigma_{n=0}^{\infty} a^n = (1-a)^{-1}$ for $0 < a < 1$. The expected value is found from

$$E[X] = \mu = (1-a)\sum_{n=0}^{\infty} na^n = (1-a) \times a \times \frac{d}{da}\{(1-a)^{-1}\} = \frac{a}{1-a}.$$

Solving for a, we obtain

$$a = \frac{\mu}{1+\mu}.$$

Thus, we can rewrite the geometric PMF as

$$P_X(n) = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^n u(n).$$

Note: There is another common definition of a geometric RV where the PMF support is $[1, \infty)$ instead of $[0, \infty)$. The corresponding geometric law appeared early in Example 1.9-4. Its PMF would take the form $P_X(n) = (1-a)a^{n-1}u(n-1)$, that is, the same sequence of numbers shifted right one place.

4.2 CONDITIONAL EXPECTATIONS

In many practical situations we want to know the *average of a subset* of the population: the average of the *passing grades* of an exam; the average lifespan of people who are still alive at age 70; the average height of fighter pilots (many air forces have both an upper and lower limit on the acceptable height of a pilot); the average blood pressure of long-distance runners, and so forth. Problems of this type fall within the realm of *conditional expectations.*

In conditional expectations we compute the average of a subset of a population that shares some property due to the outcome of an event. For example in the case of the average of passing grades, the subset is those exams that received passing grades. What all these exams share is that their grade is, say, ≥ 65. The event that has occurred is that they received passing grades.

Definition 4.2-1 The conditional expectation of X given that the event B has occurred is

$$E[X|B] \triangleq \int_{-\infty}^{\infty} x f_{X|B}(x|B)\,dx. \tag{4.2-1}$$

†That is, $u(n) = 1$ for $n \geq 0$ and $u(n) = 0$, else.

If X is discrete, then Equation 4.2-1 can be replaced with

$$E[X|B] \triangleq \sum_i x_i P_{X|B}(x_i|B). \quad \blacksquare \tag{4.2-2}$$

To give the reader a feel for the notion of conditional expectation, consider the following exam scores in a course on probability theory: 28, 35, 44, 66, 68, 75, 77, 80, 85, 87, 90, 100, 100. Assume that the passing grade is 65. Then the average score is 71.9; however, the *average passing score* is 82.8. A closely related example is worked out as follows.

Example 4.2-1

(*conditional expectation of uniform distribution*) Consider a continuous RV X and the event $B \triangleq \{X \geq a\}$. From Equations 2.6-1 and 2.6-2 and a little bit of work, we obtain

$$F_{X|B}(x|X \geq a) = \begin{cases} 0, & x < a, \\ \dfrac{F_X(x) - F_X(a)}{1 - F_X(a)}, & x \geq a. \end{cases} \tag{4.2-3}$$

Hence

$$f_{X|B}(x|X \geq a) = \begin{cases} 0, & x < a, \\ \dfrac{f_X(x)}{1 - F_X(a)}, & x \geq a \end{cases} \tag{4.2-4}$$

and

$$E[X|X \geq a] = \frac{\displaystyle\int_a^\infty x f_X(x)\,dx}{\displaystyle\int_a^\infty f_X(x)\,dx}. \tag{4.2-5}$$

Assume that X is a uniform RV in [0, 100]. Then

$$E[X] = \frac{1}{100}\int_0^{100} x\,dx = 50,$$

but using Equation 4.2-5 with $a = 65$

$$E[X|X \geq 65] = 82.5.$$

Conditional expectations often occur when dealing with RVs that are related in some way. For example let Y denote the lifetime of a person chosen at random, and let X be a binary RV that denotes whether the person smokes or not, that is, $X = 0$ if a nonsmoker, $X = 1$ if a smoker. Then clearly $E[Y|X = 0]$ is expected to be larger[†] than $E[Y|X = 1]$. Or let X be the

[†]Statistical evidence indicates that each cigarette smoked reduces longevity by about eight minutes. Hence smoking one pack a day for a whole year reduces the expected longevity of the smoker by 40 days!

intensity of the incident illumination and let Y be the instantaneous photocurrent generated by a photodetector. Typically the expected value of Y will be larger for stronger illumination and smaller for weaker illumination. We define some important concepts as follows.

Definition 4.2-2 Let X and Y be discrete RVs with joint PMF $P_{X,Y}(x_i, y_j)$. Then the conditional expectation of Y given $X = x_i$ denoted by $E[Y|X = x_i]$ is

$$E[Y|X = x_i] \triangleq \sum_j y_j P_{Y|X}(y_j|x_i). \tag{4.2-6}$$

Here $P_{Y|X}(y_j|x_i)$ is the conditional probability that $\{Y = y_j\}$ occurs given that $\{X = x_i\}$ has occurred and is given by $P_{X,Y}(x_i, y_j)/P_X(x_i)$. ■

We can derive an interesting and useful formula for $E[Y]$ in terms of the conditional expectation of Y given $X = x$. The reasoning is much the same as that which we used in computing the average or total probability of an event in terms of its conditional probabilities (see Equation 1.6-7 or 2.6-4). Thus,

$$E[Y] = \sum_j y_j P_Y(y_j) \tag{4.2-7}$$

$$= \sum_j y_j \sum_i P_{X,Y}(x_i, y_j)$$

$$= \sum_i \left[\sum_j y_j P_{Y|X}(y_j|x_i) \right] P_X(x_i)$$

$$= \sum_i E[Y|X = x_i] P_X(x_i). \tag{4.2-8}$$

Equation 4.2-8 is a very neat result and says that we can compute $E[Y]$ by averaging the conditional expectation of Y given X with respect to X.[†] Thus, in the smoking-longevity example discussed earlier, suppose $E[Y|X = 0] = 79.2$ years and $E[Y|X = 1] = 69.4$ years and $P_X(0) = 0.75$ and $P_X(1) = 0.25$. Then

$$E[Y] = 79.2 \times 0.75 + 69.4 \times 0.25 = 76.75$$

is the expected lifetime of the general population.

A result similar to Equation 4.2-8 holds for the continuous case as well. It is derived using Equation 2.6-85 from Chapter 2, that is,

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \qquad f_X(x) \neq 0. \tag{4.2-9}$$

The definition of conditional expectation for a continuous RV follows.

[†]Notice that this statement implies that the conditional expectation of Y given X is an RV. We shall elaborate on this important concept shortly. For the moment we assume that X assumes the fixed value x_i (or x).

Definition 4.2-3 Let X and Y be continuous RVs with joint pdf $f_{XY}(x,y)$. Let the conditional pdf of Y given that $X = x$ be denoted as in Equation 4.2-9. Then the conditional expectation of Y given that $X = x$ is given by

$$E[Y|X=x] \triangleq \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy. \quad \blacksquare \tag{4.2-10}$$

Since

$$E[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{XY}(x,y)\,dx\,dy, \tag{4.2-11}$$

it follows from Equations 4.2-9 and 4.2-10 that

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} f_X(x)\left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy\right] dx \\ &= \int_{-\infty}^{\infty} E[Y|X=x] f_X(x)\,dx. \end{aligned} \tag{4.2-12}$$

Equation 4.2-12 is the continuous RV equivalent of Equation 4.2-8. It can be used to good advantage (over the direct method) for computing $E[Y]$. We illustrate this point with an example from optical communications.

Example 4.2-2

(*conditional Poisson*) In the photoelectric detector shown in Figure 4.2-1, the number of photoelectrons Y produced in time τ depends on the (normalized) incident energy X. If X were constant, say $X = x$, Y would be a Poisson RV [4-4] with parameter x, but as real light

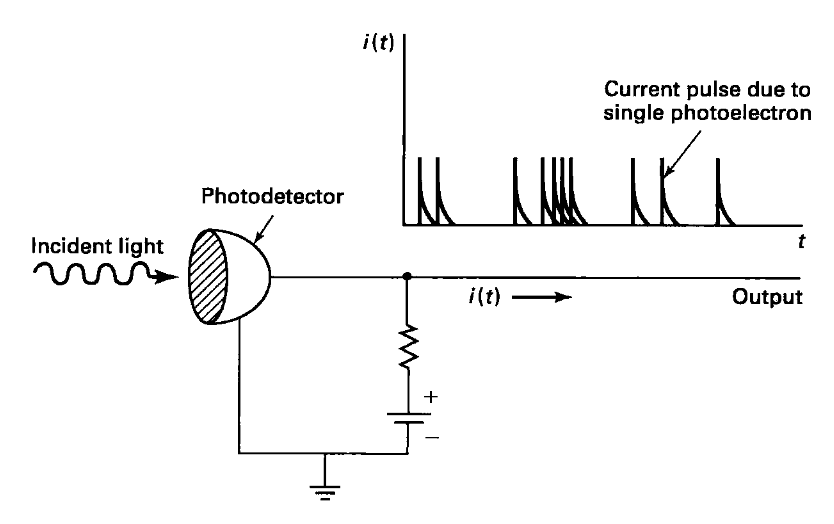


Figure 4.2-1 In a photoelectric detector, incident illumination generates a current consisting of photogenerated electrons.

sources—except for gain-stabilized lasers—do not emit constant energy signals, X must be treated as an RV. In certain situations the pdf of X is accurately modeled by

$$f_X(x) = \begin{cases} \frac{1}{\mu_X} \exp\left(-\frac{x}{\mu_X}\right), & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{4.2-13}$$

where μ_X is a parameter that equals $E[X]$. We shall now compute $E[Y]$ using Equation 4.2-12 and using the direct method.

Solution Since for $X = x$, Y is Poisson, we can write

$$P[Y = k|X = x] = \frac{x^k}{k!}e^{-x} \qquad k = 0, 1, 2, \ldots$$

and, from Example 4.1-6,

$$E[Y|X = x] = x.$$

Finally, using Equation 4.2-12 with the appropriate substitutions, that is,

$$E[Y] = \int_0^\infty x \left[\frac{1}{\mu_X} \exp\left(-\frac{x}{\mu_X}\right)\right] dx,$$

we obtain, by integration by parts,

$$E[Y] = \mu_X.$$

In contrast to the simplicity with which we obtained this result, consider the direct approach, that is,

$$E[Y] = \sum_{k=0}^{\infty} kP_Y(k). \tag{4.2-14}$$

To compute $P_Y(k)$ we use the Poisson transform (Equation 2.6-14) with $f_X(x)$, as given by Equation 4.2-13. This furnishes (see Equation 2.6-23)

$$P_Y(k) = \frac{\mu_X^k}{(1 + \mu_X)^{k+1}}. \tag{4.2-15}$$

Finally, using Equation 4.2-15 in 4.2-14 yields

$$E[Y] = \sum_{k=0}^{\infty} k \frac{\mu_X^k}{(1 + \mu_X)^{k+1}}.$$

It is known that this series sums to μ_X. Alternatively one can evaluate the sum indirectly using some clever tricks involving derivatives.

Example 4.2-3

(*conditional Gaussian density*) Let X and Y be two zero-mean RVs with joint density

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2\sigma^2(1-\rho^2)}\right) \quad |\rho|<1. \tag{4.2-16}$$

We shall soon find out (Section 4.3) that the pdf in Equation 4.2-16 is a special case of the general joint Gaussian law for two RVs. First we see that when $\rho \neq 0, f_{XY}(x,y) \neq f_X(x)f_Y(y)$; hence X and Y are not independent when $\rho \neq 0$. When $\rho = 0$, we can indeed write $f_{XY}(x,y) = f_X(x)f_Y(y)$ so that $\rho = 0$ implies independence. For the present, however, our unfamiliarity with the meaning of ρ (ρ is called the normalized *covariance* or *correlation coefficient*) is not important. When ρ is zero, X and Y are zero-mean Gaussian RVs, that is,

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}.$$

However, the conditional expectation of Y given $X = x$ is not zero even though Y is a zero-mean RV! In fact from Equation 4.2-9,

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2\sigma^2(1-\rho^2)}\right). \tag{4.2-17}$$

Hence $f_{Y|X}(y|x)$ is Gaussian with mean ρx. Thus,

$$\begin{aligned} E[Y|X=x] &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy \\ &= \rho x. \end{aligned} \tag{4.2-18}$$

When ρ is close to unity, $E[Y|X = x] \simeq x$, which implies that Y tracks X quite closely (exactly if $\rho = 1$), and if we wish to predict Y, say, with Y_P upon observing $X = x$, a good bet is to choose our *predicted value* $Y_P = x$. On the other hand, when $\rho = 0$, observing X doesn't help us to predict Y. Thus, we see that in the Gaussian case at least and somewhat more generally, ρ is related to the predictability of one RV from observing another. A cautionary note should be sounded, however: The fact that one RV doesn't help us to linearly predict another doesn't generally mean that the two RVs are independent.

Example 4.2-4

(*expectation conditioned on sums of RVs*) Consider the two independent, discrete, RVs K_1 and K_2. We wish to compute $E[K_1|K_1 + K_2 = m]$. It is first necessary to determine the conditional probability $P[K_1 = k_1|K_1 + K_2 = m]$. This conditional probability can be written as

$$\begin{aligned} P[K_1 = k_1|K_1+K_2=m] &= \frac{P[K_1=k_1, K_1+K_2=m]}{P[K_1+K_2=m]} \\ &= \frac{P[K_1=k_1, K_2=m-k_1]}{P[K_1+K_2=m]} \\ &= \frac{P[K_1=k_1]P[K_2=m-k_1]}{P[K_1+K_2=m]}. \end{aligned} \tag{4.2-19}$$

Let K_1 and K_2 be each distributed as Poisson with the same parameter θ. Since these RVs are independent and identically distributed we designate them as i.i.d. RVs. Then, from Equation 4.2-19 and Example 3.3-8 we get

$$\begin{aligned} P[K_1 = k_1|K_1 + K_2 = m] \\ &= \frac{(e^{-\theta_1}\theta_1^{k_1}/k_1!) \times (e^{-\theta_2}\theta_2^{m-k_1}/(m-k_1)!)}{e^{-(\theta_1+\theta_2)}(\theta_1+\theta_2)^m/m!} \\ &= \binom{m}{k_1}\theta_1^{k_1}\theta_2^{m-k_1} \times (\theta_1+\theta_2)^{-m}. \end{aligned} \tag{4.2-20}$$

Now recall that $E[K_1|K_1+K_2=m] \triangleq \sum_{k_1=0}^{m} k_1 P[K_1 = k_1|K_1+K_2=m]$ and the binomial expansion formula is given by $\sum_{k=0}^{n}\binom{n}{k}\theta_1^k\theta_2^{n-k} = (\theta_1+\theta_2)^n$. Then using Equation 4.2-20 finally yields

$$E[K_1|K_1+K_2=m] = m \times \left(\frac{\theta_1}{\theta_1+\theta_2}\right). \tag{4.2-21}$$

Example 4.2-5

(*continuation of Example* 4.2-4) Let K_1, K_2, K_3 denote multinomial RVs for $l = 3$, that is, a three-nomial (three outcomes possible). Then for n trials, we have the PMF†

$$\begin{aligned} P_{\mathbf{K}}(k_1,k_2,k_3) &= P[K_1 = k_1, K_2 = k_2, K_3 = k_3] \\ &= \begin{cases} \frac{n!}{k_1!k_2!k_3!}p_1^{k_1}p_2^{k_2}p_3^{k_3}, & k_1+k_2+k_3 = n, \text{ all } k_i \geq 0, \\ 0, & \text{else,} \end{cases} \end{aligned} \tag{4.2-22}$$

where $p_1 + p_2 + p_3 = 1$. We wish to compute $E[K_1|K_1 + K_2 = m]$.

Solution As in the previous example, we need to compute $P[K_1 = k_1|K_1+K_2 = m]$. We write

$$P[K_1 = k_1|K_1+K_2=m] = \frac{P[K_1 = k_1, K_1+K_2=m]}{P[K_1+K_2=m]}.$$

Note that for the multinomial, the event $\{\zeta\colon K_1(\zeta)+K_2(\zeta) = m\} \cap \{\zeta\colon K_1(\zeta) = k_1\}$ is identical to the event $\{\zeta\colon K_1(\zeta) = k_1, K_2(\zeta) = m-k_1, K_3(\zeta) = n-m\}$. Hence

†Note the notation different from that in the binomial case. Using this new multinomial notation for the binomial case, we would have, for a binomial RV K: $K_1 = K$ and $K_2 = n - K$. In the general l-nomial distribution we must always abide by the constraint that $K_1 + K_2 + ... + K_l = n$.

$$P[K_1 = k_1 | K_1 + K_2 = m] = \frac{P[K_1 = k_1, K_2 = m - k_1, K_3 = n - m]}{P[K_3 = n - m]} \tag{4.2-23}$$

$$= \frac{n!}{k_1!(m-k_1)!(n-m)!} p_1^{k_1} p_2^{m-k_1} p_3^{n-m}$$

$$\div \frac{n!}{(n-m)!m!} p_3^{n-m} (1-p_3)^m$$

$$= \binom{m}{k_1} p_1^{k_1} p_2^{m-k_1} (p_1 + p_2)^{-m} \tag{4.2-24}$$

Finally, using

$$E[K_1 | K_1 + K_2 = m] = \sum_{k_1} k_1 P[K_1 = k_1 | K_1 + K_2 = m],$$

we obtain that

$$E[K_1 | K_1 + K_2 = m] = m \frac{p_1}{p_1 + p_2}. \tag{4.2-25}$$

We leave it to the reader to compute that

$$E[K_2 | K_1 + K_2 = m] = m \frac{p_2}{p_1 + p_2}. \tag{4.2-26}$$

These kinds of problems occur in the estimation procedure known as the expectation-maximization algorithm, discussed in detail in Chapter 11.

Conditional Expectation as a Random Variable

Consider, for the sake of being specific, a function $Y = g(X)$ of a discrete RV X. Then its expected value is

$$E[Y] = \sum_i g(x_i) P_X(x_i)$$

$$= E[g(X)].$$

This suggests that we could write Equation 4.2-8 in similar notation, that is,

$$E[Y] = \sum_i E[Y | X = x_i] P_X(x_i)$$

$$= E[E[Y|X]]. \tag{4.2-27}$$

It is important to note that the object $E[Y|X = x_i]$ is a number, as is $g(x_i)$, but the object $E[Y|X]$ is a function of the RV X *and therefore is itself* an RV. Given a probability space $\mathscr{P} = (\Omega, \mathscr{F}, P)$ and an RV X defined on $\mathscr{P}$, for each outcome $\zeta \in \Omega$ we generate

the real number $E[Y|X = X(\zeta)]$. Thus, for ζ variable $E[Y|X]$ is an RV that assumes the value $E[Y|X = X(\zeta)]$ when ζ is the outcome of the underlying experiment. As always, the functional dependence of X on ζ is suppressed, and we specify X rather than the underlying probability space $\mathscr{P}$. The following example illustrates the use of the conditional expectation as an RV.

Example 4.2-6

(*multi-channel communications*) Consider a communication system in which the message delay (in milliseconds) is T and the channel choice is L. Let $L = 1$ for a satellite channel, $L = 2$ for a coaxial cable channel, $L = 3$ for a microwave surface link, and $L = 4$ for a fiber-optical link. A channel is chosen based on availability, which is a random phenomenon. Suppose $P_L(l) = 1/4$, $l = 1, \ldots, 4$. Assume that it is known that $E[T|L = 1] = 500$, $E[T|L = 2] = 300$, $E[T|L = 3] = 200$, and $E[T|L = 4] = 100$. Then the RV $g(L) \triangleq E[T|L]$ is defined by

$$g(L) = \begin{cases} 500, & \text{for } L = 1 \quad P_L(1) = \frac{1}{4}, \\ 300, & \text{for } L = 2 \quad P_L(2) = \frac{1}{4}, \\ 200, & \text{for } L = 3 \quad P_L(3) = \frac{1}{4}, \\ 100, & \text{for } L = 4 \quad P_L(4) = \frac{1}{4}, \end{cases}$$

and $E[T] = E[g(L)] = 500 \times \frac{1}{4} + 300 \times \frac{1}{4} + 200 \times \frac{1}{4} + 100 \times \frac{1}{4} = 275$.

The notion of $E[Y|X]$ being an RV is equally valid for discrete, continuous, or mixed RVs X. For example, Equation 4.2-12

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x] f_X(x)\, dx$$

can also be written as $E[Y] = E\left[E[Y|X]\right]$, where $E[Y|X]$ in this case is a function of the continuous RV X. The inner expectation is with respect to Y and the outer with respect to X.

The foregoing can be extended to more complex situations. For example, the object $E[Z|X, Y]$ is a function of the RVs X and Y and therefore is a function of two RVs. For a particular outcome $\zeta \in \Omega$, it assumes the value $E[Z|X(\zeta), Y(\zeta)]$. To compute $E[Z]$ we would write $E[Z] = E[E[Z|X, Y]]$, which, for example, in the case of continuous RVs yields

$$\begin{aligned} E[Z] &= E[E[Z|X, Y]] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z f_{Z|X,Y}(z|x, y)\, f_{XY}(x, y) dx dy dz. \end{aligned} \tag{4.2-28}$$

We conclude this section by summarizing some properties of conditional expectations.

Properties of Conditional Expectation:.

Property (i). $E[Y] = E[E[Y|X]]$.

Proof See arguments leading up to Equation 4.2-8 for the discrete case and Equation 4.2-12 for the continuous case. The inner expectation is with respect to Y, the outer with respect to X.

Property (ii). If X and Y are independent, then $E[Y|X] = E[Y]$.

Proof

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy.$$

But $f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x) = f_Y(y)f_X(x)$ if X and Y are independent. Hence $f_{Y|X}(y|x) = f_Y(y)$ and

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_Y(y)dy = E[Y]$$

for each x. Thus,

$$E[Y|X] = \int_{-\infty}^{\infty} y f_Y(y)dy = E[Y].$$

An analogous proof holds for the discrete case.

Property (iii). $E[Z|X] = E[E[Z|X, Y]|X]$.

Proof

$$\begin{aligned}
E[Z|X = x] &= \int_{-\infty}^{\infty} z f_{Z|X}(z|x)dz \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z f_{Z|X,Y}(z|x, y) f_{Y|X}(y|x)dz\, dy \\
&= \int_{-\infty}^{\infty} dy\, f_{Y|X}(y|x) \int_{-\infty}^{\infty} z f_{Z|X,Y}(z|x, y)dz \\
&= E\left[E[Z|X, Y]|X = x\right],
\end{aligned}$$

where the inner expectation is with respect to Z and the outer with respect to Y. Since this is true for all x, we have $E[Z|X] = E[E[Z|X, Y]|X]$. The mean $\mu_Y = E[Y]$ is an estimate of the RV Y. The mean-square error in this estimate is $\varepsilon^2 = E[(Y - \mu_Y)^2]$. In fact this estimate is optimal in that any constant other than μ_Y would lead to an increased ε^2. ■

4.3 MOMENTS OF RANDOM VARIABLES

Although the expectation is an important "summary" number for the behavior of an RV, it is far from adequate in describing the complete behavior of the RV. Indeed, we saw in Section 4.1 that two sets of numbers could have the same sample mean but the sample deviations could be quite different. Likewise, for two RVs their expectations could be the same but their standard deviations could be very different. Summary numbers like μ_X, σ_X^2, $E[X^2]$, and others are called *moments*. Generally, an RV will have many nonzero higher-order moments and, under certain conditions (Section 4.5), it is possible to completely describe the behavior of the RV, that is, reconstruct its pdf from knowledge of all the moments. In the following definitions we shall assume that the moments exist. However, this is not always the case.

Definition 4.3-1 The rth moment of X is defined as

$$m_r \triangleq E[X^r] = \int_{-\infty}^{\infty} x^r f_X(x)\, dx, \quad \text{where } r = 0, 1, 2, 3, \ldots. \tag{4.3-1}$$

If X is a discrete RV, the rth moment can be computed from the PMF as

$$m_r \triangleq \sum_i x_i^r P_X(x_i).$$

We note that $m_0 = 1$, $m_1 = \mu$ (the mean). ■

Definition 4.3-2 The rth *central moment* of X is defined as

$$c_r \triangleq E[(X - \mu)^r], \qquad \text{where } r = 0, 1, 2, 3, \ldots. \tag{4.3-2a}$$

For a discrete RV we can compute c_r from

$$c_r \triangleq \sum_i (x_i - \mu)^r P_X(x_i). \quad ■ \tag{4.3-2b}$$

The most frequently used central moment is c_2. It is called the *variance* and is denoted by σ^2 and also sometimes by Var$[X]$. Note that $c_0 = 1$, $c_1 = 0$, $c_2 = \sigma^2$. An important formula that connects the variance to $E[X^2]$ and μ is obtained as follows:

$$\sigma^2 = E\left[[X - \mu]^2\right] = E[X^2] - E[2\mu X] + E[\mu^2].$$

But for any constant a, $E[aX] = aE[X]$ and $E[a^2] = a^2$. Thus

$$\begin{aligned} \sigma^2 &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned} \tag{4.3-3}$$

since $E[X] \triangleq \mu$. In order to save symbology, an overbar is often used to denote expectation. Thus $\overline{X^r} \triangleq E[X^r]$, and so forth, for other moments. Using this notation, Equation 4.3-3 appears as

$$\sigma^2 = \overline{X^2} - \mu^2 \tag{4.3-4a}$$

or, equivalently,

$$\overline{X^2} = \sigma^2 + \mu^2. \tag{4.3-4b}$$

Equations 4.3-4a relates the second central moment c_2 to μ_2 and μ. We can generalize this result as follows. Observe that

$$(X-\mu)^r = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \mu^i X^{r-i}. \tag{4.3-5a}$$

By taking the expectation of both sides of Equation 4.3-5a and recalling the linearity of the expectation operator, we obtain

$$c_r = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \mu^i m_{r-i}. \tag{4.3-5b}$$

Example 4.3-1

Let us compute m_2 for X, a binomial RV. By definition

$$P_X(k) = \binom{n}{k} p^k q^{n-k}$$

and

$$\begin{aligned} m_2 &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k} \\ &= p^2 n(n-1) + np \\ &= n^2p^2 + npq. \end{aligned} \tag{4.3-6}$$

In going from line 2 to line 3 several steps of algebra were used whose duplication we leave as an exercise. In going from line 3 to line 4, we rearranged terms and used the fact that $q \triangleq 1-p$. The expected value of X is

$$\begin{aligned} m_1 &= \sum_{k=0}^{n} k \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &= np = \mu. \end{aligned} \tag{4.3-7}$$

Using this result in Equation 4.3-6 and recalling Equation 4.3-4 allow us to conclude that for a binomial RV with PMF $b(k; n, p)$

$$\sigma^2 = npq. \tag{4.3-8}$$

For any given n, maximum variance is obtained when $p = q = 0.5$ (Figure 4.3-1).

Example 4.3-2

(*second moment of zero-mean Gaussian*) Let us compute central moment c_2 for $X : N(0, \sigma^2)$. Since $\mu = 0$, $c_2 = m_2$ and

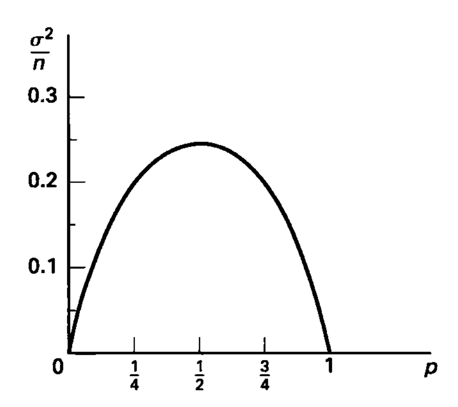


Figure 4.3-1 Variance of a binomial RV versus p.

$$c_2 = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}(x/\sigma)^2}\, dx.$$

But this integral was already evaluated in Example 4.1-7, where we found $E[X^2] = \sigma^2$. Thus, the variance of a Gaussian RV is indeed the parameter σ^2 regardless of whether X is zero-mean or not.

An interesting and somewhat more difficult example that illustrates a useful application of moments is given next.

Example 4.3-3

(*entropy*) The *maximum entropy* (ME) principle states that if we don't know the pdf $f_X(x)$ of X but would like to estimate it with a function, say $p(x)$, a good choice is the function $p(x)$ which maximizes the *entropy*, defined by [4-5],

$$H[X] \triangleq -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx \tag{4.3-9}$$

and which satisfies the constraints

$$p(x) \geq 0 \tag{4.3-10a}$$

$$\int_{-\infty}^{\infty} p(x)\, dx = 1 \tag{4.3-10b}$$

$$\int_{-\infty}^{\infty} xp(x)\, dx = \mu \tag{4.3-10c}$$

$$\int_{-\infty}^{\infty} x^2 p(x)\, dx = m_2, \text{ and so forth.} \tag{4.3-10d}$$

Suppose we know from measurements or otherwise only μ in Equation 4.3-10c and that $x \geq 0$. Thus, we wish to find $p(x)$ that maximizes $H[X]$ of Equation 4.3-9 subject to

the first three constraints of Equation 4.3-10. Using the method of Lagrange multipliers [4-6], the solution is obtained by maximizing the expression

$$-\int_0^\infty p(x)\ln p(x)\,dx - \lambda_1\int_0^\infty p(x)\,dx - \lambda_2\int_0^\infty xp(x)\,dx$$

by differentiation with respect to $p(x)$. The constants λ_1 and λ_2 are Lagrange multipliers and must be determined. After differentiating we obtain

$$\ln p(x) = -(1+\lambda_1) - \lambda_2 x$$

or

$$p(x) = e^{-(1+\lambda_1+\lambda_2 x)}. \tag{4.3-11}$$

When this result is substituted in Equations 4.3-10b and 4.3-10c, we find that

$$e^{-(1+\lambda_1)} = \frac{1}{\mu}, \quad \mu > 0,$$

and

$$\lambda_2 = \frac{1}{\mu}.$$

Hence our ME estimate of $f_X(x)$ is

$$p(x) = \begin{cases} \frac{1}{\mu}e^{-x/\mu}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{4.3-12}$$

The problem of obtaining the ME estimate of $f_X(x)$ when both μ and σ^2 are known is left as an exercise. In this case $p(x)$ is the Normal distribution with mean μ and variance σ^2.

Tables of common means, variances, and mean-square values. Table 4.3-1 is a table of means, variances, and mean-square values for common continuous RVs. Some of these have been calculated already in the text. Others are left as end-of-chapter problems.

Table 4.3-2 is a similar table for common discrete RVs.

Less useful than m_r or c_r are the absolute moments and generalized moments about some arbitrary point, say a, defined by, respectively,

$$E[|X|^r] \triangleq \int_{-\infty}^{\infty} |x|^r f_X(x)\,dx \quad \text{(absolute moment)}$$

$$E[(X-a)^r] \triangleq \int_{-\infty}^{\infty} (x-a)^r f_X(x)\,dx \quad \text{(generalized moment)}.$$

Note that if we set $a = \mu$, the generalized moments about a are then the central moments. If $a = 0$, the generalized moments are simply the moments m_r.

Table 4.3-1 Means, Variances and Mean-Square values for Common Continuous RVs

Family	pdf $f(x)$	Mean $\mu = E[X]$	Variance σ^2	Mean square $E[X^2]$
Uniform	$U(a,b)$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$	$\frac{1}{3}(b^2+ab+a^2)$
Exponential	$\frac{1}{\mu}e^{-x/\mu}u(x)$	μ	μ^2	$2\mu^2$
Gaussian	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$\mu^2+\sigma^2$
Laplacian	$\frac{1}{\sqrt{2}\sigma}e^{-\frac{\sqrt{2}}{\sigma}\lvert x\rvert}$	0	σ^2	σ^2
Rayleigh	$\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}u(x)$	$\sqrt{\frac{\pi}{2}}\sigma$	$\left(2-\frac{\pi}{2}\right)\sigma^2$	$2\sigma^2$

Table 4.3-2 Means, Variances, and Mean-Square Values for Common Discrete RVs

Family	PMF $P(k)$	Mean $\mu = E[K]$	Variance σ^2	Mean square $E[K^2]$
Bernoulli	$P_B(k) = \begin{cases} 1, & p \\ 0, & q \end{cases}\ q \triangleq 1-p$	p	pq	p
Binomial	$b(k;n,p) = \binom{n}{k}p^k q^{n-k}$	np	npq	$(np)^2+npq$
Geometric†	$\frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^k u(k)$	μ	$\mu+\mu^2$	$\mu+2\mu^2$
Poisson	$\frac{\alpha^k}{k!}e^{-\alpha}u(k)$	α	α	$\alpha^2+\alpha$

Joint Moments

Let us now turn to a topic first touched upon in Example 4.2-3. Suppose we are given two RVs X and Y and wish to have a measure of how good a linear prediction we can make of the value of, say, Y upon observing what value X has. At one extreme if X and Y are independent, observing X tells us nothing about Y. At the other extreme if, say, $Y = aX + b$, then observing the value of X immediately tells us the value of Y. However, in many situations in the real world, two RVs are neither completely independent nor linearly dependent. Given this state of affairs, it then becomes important to have a measure of how much can be said about one RV from observing another. The quantities called *joint moments* offer us such a measure. Not all joint moments, to be sure, are equally important in this task; especially important are certain *second-order joint moments* (to be defined shortly). However, as we shall see later, in various applications other joint moments are important as well and so we shall deal with the general case below.

† The geometric PMF is sometimes written in terms of the parameter a as $(1-a)a^k u(k)$ with $0 < a < 1$. Then $\mu = a/(1-a)$ with $\mu > 0$.

Definition 4.3-3 The ijth joint moment of X and Y is given by

$$m_{ij} \triangleq E[X^i Y^j]$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^i y^j f_{XY}(x,y)\,dx\,dy. \quad (4.3\text{-}13)$$

If X and Y are discrete, we can compute μ_{ij} from the PMF as

$$m_{ij} \triangleq \sum_l \sum_m x_l^i y_m^j P_{X,Y}(x_l, y_m). \quad \blacksquare \quad (4.3\text{-}14)$$

Definition 4.3-4 The ijth joint *central moment* of X and Y is defined by

$$c_{ij} \triangleq E[(X-\overline{X})^i (Y-\overline{Y})^j], \quad (4.3\text{-}15)$$

where, in the notation introduced earlier, $\overline{X} \triangleq E[X]$, and so forth, for $\overline{Y}$. The order of the moment is $i+j$. Thus, all of the following are second-order moments:

$$m_{02} = E[Y^2] \qquad c_{02} = E[(Y-\overline{Y})^2]$$
$$m_{20} = E[X^2] \qquad c_{20} = E[(X-\overline{X})^2]$$
$$m_{11} = E[XY] \qquad c_{11} = E[(X-\overline{X})(Y-\overline{Y})]$$
$$= E[XY] - \overline{X}\,\overline{Y}$$
$$\triangleq \text{Cov}[X,Y]. \quad \blacksquare$$

As measures of predictability and in some cases statistical dependence, the most important joint moments are m_{11} and c_{11}; they are known as the *correlation* and *covariance* of X and Y, respectively. The *correlation coefficient*† defined by

$$\rho \triangleq \frac{c_{11}}{\sqrt{c_{02}c_{20}}} \quad (4.3\text{-}16)$$

was already introduced in Section 4.2 (Equation 4.2-16). It satisfies $|\rho| \leq 1$. To show this consider the nonnegative expression

$$E[(\lambda(X-\mu_X) - (Y-\mu_Y))^2] \geq 0,$$

where λ is any real constant. To verify that the left side is indeed nonnegative, we merely rewrite it in the form

$$Q(\lambda) \triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\lambda(x-\mu_X) - (y-\mu_Y)]^2 f_{XY}(x,y)\,dx\,dy \geq 0,$$

†Note that it would be more properly termed the *covariance coefficient* or *normalized covariance*.

where the $\geq$ follows from the fact that the integral of a nonnegative quantity cannot be negative.

The previous equation is a quadratic in λ. Indeed, after expanding we obtain

$$Q(\lambda) = \lambda^2 c_{20} + c_{02} - 2\lambda c_{11} \geq 0.$$

Thus $Q(\lambda)$ can have at most one real root. Hence its discriminant must satisfy

$$\left(\frac{c_{11}}{c_{20}}\right)^2 - \frac{c_{02}}{c_{20}} \leq 0$$

or

$$c_{11}^2 \leq c_{02}c_{20} \tag{4.3-17}$$

whence the condition $|\rho| \leq 1$ follows.

When $c_{11}^2 = c_{02}c_{20}$, that is, $|\rho| = 1$, it is readily established that

$$E\left[\left(\frac{c_{11}}{c_{20}}(X - \mu_X) - (Y - \mu_Y)\right)^2\right] = 0$$

or, equivalently, that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{c_{11}}{c_{20}}(x - \mu_X) - (y - \mu_Y)\right)^2 f_{XY}(x, y)\,dx\,dy = 0. \tag{4.3-18}$$

Since $f_{XY}(x, y)$ is never negative, Equation 4.3-18 implies that the term in parentheses is zero everywhere.† Thus, we have from Equation 4.3-18 that when $|\rho| = 1$

$$Y = \frac{c_{11}}{c_{20}}(X - \mu_X) + \mu_Y, \tag{4.3-19}$$

that is, Y is a linear function of X. When $\mathrm{Cov}[X, Y] = 0$, $\rho = 0$ and X and Y are said to be *uncorrelated.*

Properties of Uncorrelated Random Variables

(a) If X and Y are uncorrelated, then

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2, \tag{4.3-20}$$

where

$$\sigma_{X+Y}^2 \triangleq E[(X + Y)^2] - (E[X + Y])^2.$$

†Except possibly over a bizarre set of points of zero probability. To be more precise, we should exchange the word "everywhere" in the text to "almost everywhere," often abbreviated a.e.

(b) If X and Y are independent, they are uncorrelated. Proof of (a): We leave this as an exercise to the reader; proof of (b): Since $\text{Cov}[X,Y] = E[XY] - E[X]E[Y]$, we must show that $E[XY] = E[X]E[Y]$. But

$$
\begin{aligned}
E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} x f_X(x)\,dx \int_{-\infty}^{\infty} y f_Y(y)\,dy \quad \text{(by independence assumption)} \\
&= E[X]E[Y]. \quad \blacksquare
\end{aligned}
$$

Example 4.3-4

(*linear prediction*) Suppose we wish to predict the values of an RV Y by observing the values of another RV X. In particular, the available data (Figure 4.3-2) suggest that a good prediction model for Y is the linear function

$$Y_P \triangleq \alpha X + \beta. \tag{4.3-21}$$

Now although Y may be related to X, the values it takes on may be influenced by other sources that do not affect X. Thus, in general, $|\rho| \neq 1$ and we expect that there will be an error between the predicted value of Y, that is, Y_P, and the value that Y actually assumes. Our task becomes then to adjust the coefficients α and β in order to minimize the mean-square error

$$\varepsilon^2 \triangleq E[(Y - Y_P)^2]. \tag{4.3-22}$$

This problem is a simple version of *optimum linear prediction.* In statistics it is called *linear regression.*

Solution Upon expanding Equation 4.3-22, we obtain

$$\varepsilon^2 = E[Y^2] - 2\alpha\mu_{XY} - 2\beta\mu_Y + 2\alpha\beta\mu_X + \alpha^2 E[X^2] + \beta^2.$$

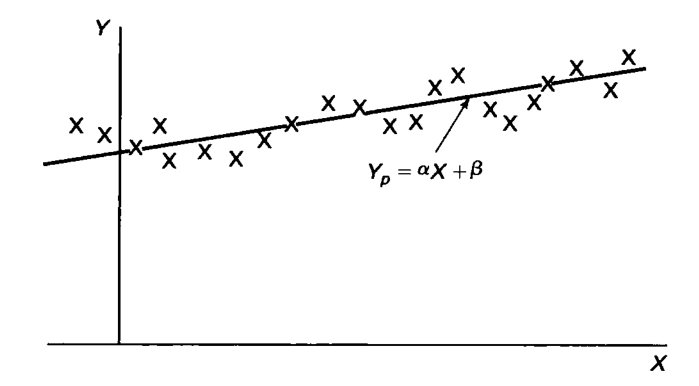


Figure 4.3-2 Pairwise observations on (X, Y) constitute a scatter diagram. The relationship between X and Y is approximated with a straight line.

To minimize ε with respect to α and β, we solve for α and β that satisfy

$$\frac{\partial \varepsilon^2}{\partial \alpha} = 0 \qquad \frac{\partial \varepsilon^2}{\partial \beta} = 0. \tag{4.3-23}$$

This yields the best α and β, which we denote by α_0, β_0 in the sense that they minimize ε. A little algebra establishes that

$$\alpha_0 = \frac{\text{Cov}[X,Y]}{\sigma_X^2} = \frac{\rho\sigma_Y}{\sigma_X} \tag{4.3-24a}$$

and

$$\begin{aligned}\beta_0 &= \overline{Y} - \frac{Cov[X,Y]}{\sigma_X^2}\overline{X} \\ &= \overline{Y} - \rho\frac{\sigma_Y}{\sigma_X}\overline{X}.\end{aligned} \tag{4.3-24b}$$

Thus, the best linear predictor is given by

$$Y_P - \mu_Y = \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \tag{4.3-25}$$

and passes through the point (μ_X, μ_Y). If we use α_0, β_0 in Equation 4.3-22 we obtain the smallest mean-square error $\varepsilon^2_{\text{min}}$, which is Problem 4.33,

$$\varepsilon^2_{\text{min}} = \sigma_Y^2(1-\rho^2). \tag{4.3-26}$$

Something rather strange happens when $\rho = 0$. From Equation 4.3-25 we see that for $\rho = 0$, $Y_P = \mu_Y$ regardless of X! This means that observing X has no bearing on our prediction of Y, and the best predictor is merely $Y_P = \mu_Y$. We encountered somewhat the same situation in Example 4.2-3. Thus, associating the correlation coefficient with ability to predict seems justified in problems involving linear prediction and the joint Gaussian pdf. In some fields, a lack of correlation between two RVs is taken to be *prima facie* evidence that they are unrelated, that is, independent. No doubt this conclusion arises in part from the fact that if two RVs, say, X and Y, are indeed independent, they will be uncorrelated. As stated earlier, the opposite is generally not true. An example follows.

Example 4.3-5

(*uncorrelated is weaker than independence*) Consider two RVs X and Y with joint PMF $P_{X,Y}(x_i, y_j)$ as shown.

Values of $P_{X,Y}(x_i, y_j)$

	$x_1 = -1$	$x_2 = 0$	$x_3 = +1$
$y_1 = 0$	0	$\frac{1}{3}$	0
$y_2 = 1$	$\frac{1}{3}$	0	$\frac{1}{3}$

X and Y are not independent, since $P_{XY}(0,1) = 0 \neq P_X(0)P_Y(1) = \frac{2}{9}$. Furthermore, $\mu_X = 0$ so that $\text{Cov}(X,Y) = E[XY] - \mu_X \mu_Y = E[XY]$. We readily compute

$$m_{11} = (-1)(1)\tfrac{1}{3} + (1)(1)\tfrac{1}{3} = 0.$$

Hence X and Y are uncorrelated but not independent.

There is an important special case for which $\rho = 0$ always implies independence. We now discuss this case.

Jointly Gaussian Random Variables

We say that two RVs are jointly Gaussian† (or jointly Normal) if their joint pdf is

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left(\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right). \tag{4.3-27}$$

Five parameters are involved: σ_X, σ_Y, μ_X, μ_Y, and ρ. If $\rho = 0$ we observe that

$$f_{XY}(x,y) = f_X(x)f_Y(y),$$

where

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \tag{4.3-28}$$

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right). \tag{4.3-29}$$

Thus, two jointly Gaussian RVs that are uncorrelated (i.e., $\rho = 0$) are also independent. The marginal densities $f_X(x)$ and $f_Y(y)$ for jointly normal RVs are always normal regardless of what ρ is. However, the converse does not hold; that is, if $f_X(x)$ and $f_Y(y)$ are Gaussian, one cannot conclude that X and Y are jointly Gaussian.

To see this we borrow from a popular x-ray imaging technique called *computerized tomography* (CT) useful for detecting cancer and other abnormalities in the body. Suppose we have an object with x-ray absorptivity function $f(x,y) \geq 0$. This function is like a joint pdf in that it is real, never negative, and easily normalized to a unit volume—however, this last feature is not important. Thus, we can establish a one-to-one relationship between a joint pdf $f_{XY}(x,y)$ and the x-ray absorptivity $f(x,y)$. In CT, x-rays are passed through the

†The jointly Normal pdf is sometimes called the two-dimensional Normal pdf in anticipation of the general multi-dimensional Normal pdf. The later becomes very cumbersome to write without using matrix notation (Chapter 5).

object along different lines, for some fixed angle, and the integrals of the absorptivity are measured and recorded. Each integral is called a projection and the set of all projections for given angle θ is called the profile function at θ. Thus, the projection for a line at angle θ and displacement s from the center is given by [Figure. 4.3-3(a)]

$$f_\theta(s) = \int_{L(s,\theta)} f(x,y)dl,$$

where $L(s,\theta)$ are the points along a line displaced from the center by s at angle θ and dl is a differential length along $L(s,\theta)$. If we let s vary from its smallest to largest value, we obtain the profile function for that angle. By collecting all the profiles for all the angles and using a sophisticated algorithm called *filtered-convolution back-projection*, it is possible to get a high-quality x-ray image of the body. Suppose we measure the profiles at 0 degrees and 90 degrees as shown in Figure 4.3-3(b). Then we obtain

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad \text{(horizontal profile)}$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \qquad \text{(vertical profile).}$$

If $f(x,y)$ is Gaussian, then we already know that $f_1(x)$ and $f_2(y)$ will be Gaussian because f_1 and f_2 are analogous to marginal pdfs. Now is it possible to modify $f(x,y)$ from Gaussian to non-Gaussian without observing a change in the Gaussian profile? If yes, we have demonstrated our assertion that Gaussian marginals do not necessarily imply a joint Gaussian pdf. In Figure 4.3-3(c) we increase the absorptivity of the object by an amount P along the 45-degree strip running from a to b and decrease the absorptivity by the same amount P along the 135-degree strip running from a' to b'. Then since the profile integrals add and

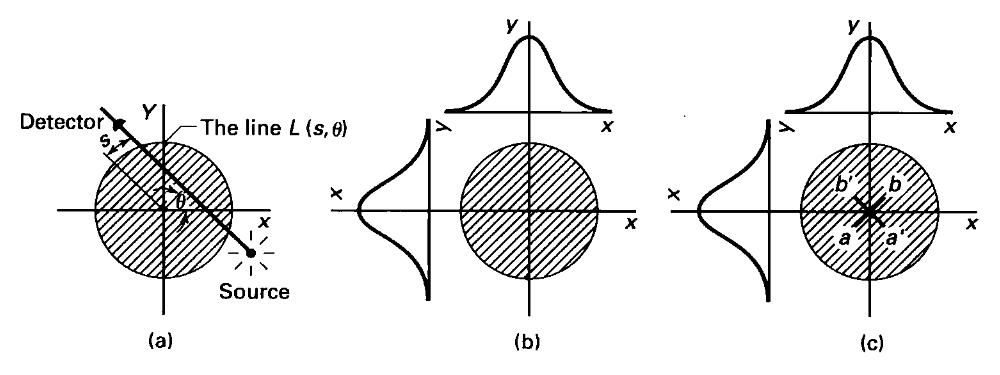


Figure 4.3-3 Using the computerized tomography paradigm to show that Gaussian marginal pdf's do not imply a joint Gaussian distribution. (a) A projection is the line integral at displacement s and angle θ. The set of all projections for a given angle is the profile function for that angle. (b) A joint Gaussian x-ray object produces Gaussian-shaped profile functions in the horizontal and vertical directions; (c) by adding a constant absorptivity along a–b and subtracting an absorptivity along a'–b', the profile functions remain the same but the underlying absorptivity is not Gaussian anymore.

subtract P in both horizontal and vertical directions, the net change in $f_1(x)$ and $f_2(y)$ is zero. This proves our assertion. We assume that P is not so large that when subtracted from $f(x,y)$ along a'–b', the result is negative. The reason we must make this assumption is that pdf's and x-ray absorptivities can never be negative.

To illustrate a joint normal distribution consider the following somewhat idealized situation. Let X and Y denote the height of the husband and wife, respectively, of a married pair picked at random from the population of married people. It is often assumed that X and Y are individually Gaussian although this is obviously only an approximation since heights are bounded from below by zero and from above by physiological constraints. Conventional wisdom has it that in our society tall people prefer tall mates and short people prefer short mates. If this is indeed true, then X and Y are *positively correlated*, that is, $\rho > 0$. On the other hand, in certain societies it may be fashionable for tall men to marry short women and for tall women to marry short men. Again we can expect X and Y to be correlated albeit negatively this time, that is, $\rho < 0$. Finally, if all marriages are the result of a lottery, we would expect ρ to be zero or very small.[†]

***Contours of constant density of the joint Gaussian pdf.** It is of interest to determine the locus of points in the xy plane when $f_{XY}(x,y)$ is set constant. Clearly $f_{XY}(x,y)$ will be constant if the exponent is set to a constant, say, a^2:

$$\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 = a^2.$$

This is the equation of an ellipse centered at $x = \mu_X$, $y = \mu_Y$. For simplicity we set $\mu_X = \mu_Y = 0$. When $\rho = 0$, the major and minor diameters of the ellipse are parallel to the x- and y-axes, a condition we know to associate with independence of X and Y. If $\rho = 0$ and $\sigma_X = \sigma_Y$, the ellipse degenerates into a circle. Several cases are shown in Figure 4.3-4.

Surprisingly the marginal densities $f_X(x)$ and $f_Y(y)$ computed from the joint pdf of Equation 4.3-27 do not depend on the parameter ρ. To see this we compute

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$$

with $\mu_X = \mu_Y = 0$ for simplicity. The integration, while somewhat messy, is easily done by following these three steps:

1. Factor out of the integral all terms that do not depend on y;
2. Complete the squares in the exponent of e (see "completing the square" in Appendix A); and
3. Recall that for $b > 0$ and real y

$$\frac{1}{\sqrt{2\pi b^2}}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{y-a}{b}\right)^2\right] dy = 1.$$

[†]In statistics it is quite difficult to observe zero correlation between two random variables, even when in theory they would be expected to be uncorrelated. The phenomenon of small, random correlations is used by hucksters and others to prove a point, which in reality is not valid.

*Starred material can be omitted on a first reading.

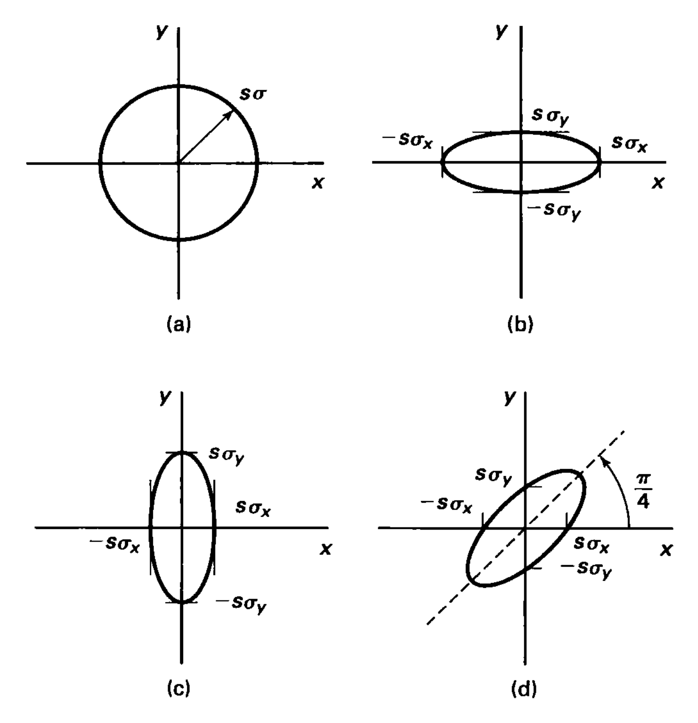


Figure 4.3-4 Contours of constant density for the joint normal ($\overline{X} = \overline{Y} = 0$): (a) $\sigma_X = \sigma_Y$, $\rho = 0$; (b) $\sigma_X > \sigma_Y$, $\rho = 0$; (c) $\sigma_X < \sigma_Y$, $\rho = 0$; (d) $\sigma_X = \sigma_Y$; $\rho > 0$.

Indeed after step 2 we obtain

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma_X}\right)^2\right]$$
$$\times \left\{\frac{1}{\sqrt{2\pi\sigma_Y^2(1-\rho^2)}} \int_{-\infty}^{\infty} \exp\left[\frac{-(y-\rho x\sigma_Y/\sigma_X)^2}{2\sigma_Y^2(1-\rho^2)}\right] dy\right\}. \tag{4.3-30}$$

But the term in curly brackets is unity. Hence

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma_X}\right)^2\right]. \tag{4.3-31}$$

A similar calculation for $f_Y(y)$ would furnish

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_Y}\right)^2\right]. \tag{4.3-32}$$

As we stated earlier, if $\rho = 0$, then X and Y are independent. On the other hand as $\rho \to \pm 1$, X and Y become linearly dependent. For simplicity let $\sigma_X = \sigma_Y \triangleq \sigma$ and $\mu_X = \mu_Y = 0$; then the contour of constant density becomes

$$x^2 - 2\rho xy + y^2 = c^2\sigma^2,$$

which is a 45-degree tilted ellipse (with respect to the x-axis) for $\rho > 0$ and a 135-degree tilted ellipse for $\rho < 0$. We can generate a coordinate system that is rotated 45 -degrees from the $x - y$ system by introducing the coordinate transformation

$$v = \frac{x+y}{\sqrt{2}} \quad w = \frac{x-y}{\sqrt{2}}.$$

Then the contour of constant density becomes

$$v^2[1-\rho] + w^2[1+\rho] = \sigma^2 c^2,$$

which is an ellipse with major and minor diameters parallel to the v and w-axes. If $\rho > 0$, the major diameter is parallel to the v-axis; if $\rho < 0$, the major diameter is parallel to the w-axis. As $\rho \to \pm 1$, the lengths of the major diameters become infinitely long and all of the pdf concentrates along the line $y = x(\rho \to 1)$ or $y = -x(\rho \to -1)$.

Finally by introducing two new RVs

$$V \triangleq (X+Y)/\sqrt{2}$$

$$W \triangleq (X-Y)/\sqrt{2},$$

we find that as $\rho \to 1$

$$f_{XY}(x,y) \to \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right] \times \delta(y-x)$$

or, equivalently,

$$f_{XY}(x,y) \to \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2\right] \times \delta(y-x).$$

This degeneration of the joint Gaussian into a pdf of only one variable along the line $y = x$ is due to the fact that as $\rho \to 1$, X and Y become equal. We leave the details as an exercise to the student.

A computer rendition of the joint Gaussian pdf and its contours of constant density is shown in Figure 4.3-5 for $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_X = 2$, and $\rho = 0.9$.

4.4 CHEBYSHEV AND SCHWARZ INEQUALITIES

The Chebyshev[†] inequality furnishes a bound on the probability of how much an RV X can deviate from its mean value μ_X.

Theorem 4.4-1 (*Chebyshev inequality*) Let X be an arbitrary RV with mean μ_X and finite variance σ^2. Then for any $\delta > 0$

$$P[|X - \mu_X| \geq \delta] \leq \frac{\sigma^2}{\delta^2}. \tag{4.4-1}$$

[†]Pafnuty L. Chebyshev (1821–1894), Russian mathematician.

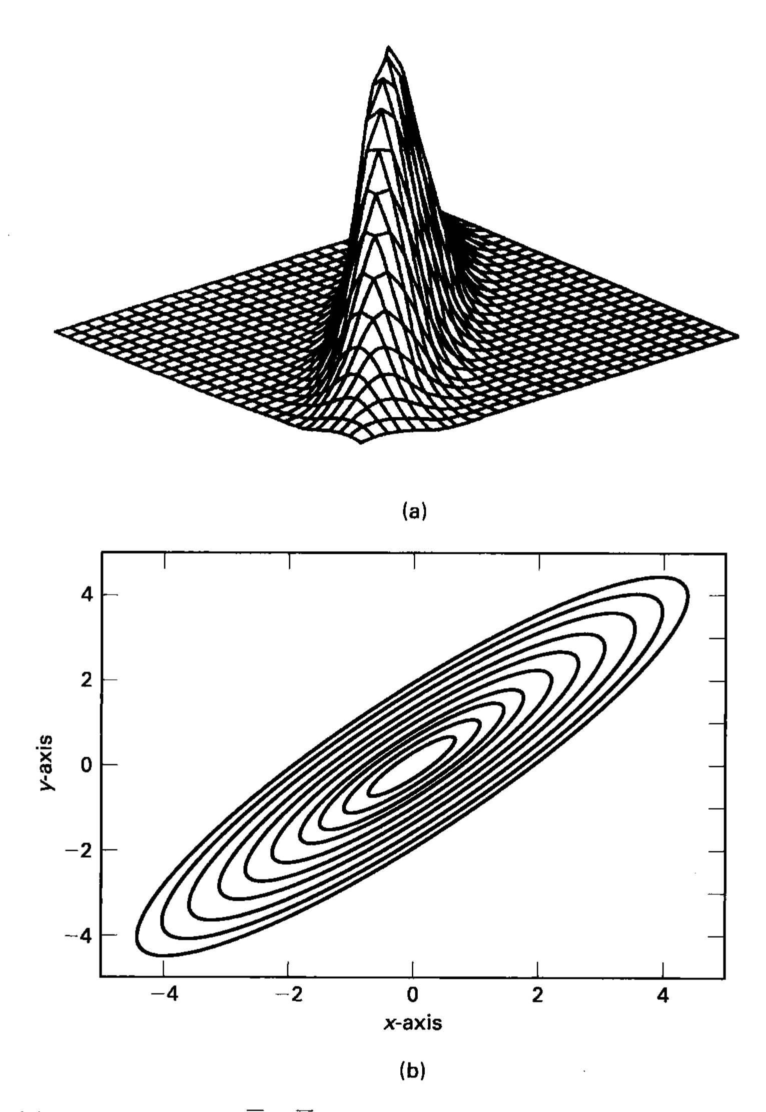


Figure 4.3-5 (a) Gaussian pdf with $\overline{X} = \overline{Y} = 0$, $\sigma_X = \sigma_Y = 2$, and $\rho = 0.9$; (b) contours of constant density.

Proof Equation 4.4-1 follows directly from the following observation:

$$\sigma^2 \triangleq \int_{-\infty}^{\infty} (x - \overline{X})^2 f_X(x)\, dx \geq \int_{|x-\overline{X}|\geq\delta} (x - \overline{X})^2 f_X(x)\, dx$$

$$\geq \delta^2 \int_{|x-\overline{X}|\geq\delta} f_X(x)\, dx$$

$$= \delta^2 P[|X - \overline{X}| \geq \delta].$$

Since

$$\{|X - \overline{X}| \geq \delta\} \cup \{|X - \overline{X}| < \delta\} = \Omega \quad (\Omega \text{ being the certain event}),$$

and the two events being unioned are disjoint, it follows that

$$P[|X - \overline{X}| < \delta] \geq 1 - \frac{\sigma^2}{\delta^2}. \tag{4.4-2}$$

Sometimes it is convenient to express δ in terms of σ, that is, $\delta \triangleq k\sigma$, where k is a constant.† Then Equations 4.4-1 and 4.4-2 become, respectively,

$$P[|X - \overline{X}| \geq k\sigma] \leq \frac{1}{k^2} \tag{4.4-3}$$

$$P[|X - \overline{X}| < k\sigma] \geq 1 - \frac{1}{k^2}. \quad \blacksquare \tag{4.4-4}$$

Example 4.4-1

(*deviation from the mean for a Normal RV*) Let $X\colon N(\mu_X, \sigma^2)$. How do $P[|X - \mu_X| < k\sigma]$ and $P[|X - \mu_X| \geq k\sigma]$ compare with the Chebyshev bound (CB)?

Solution Using Equations 2.4-14d and 2.4-14e, it is easy to show that $P[|X - \mu_X| < k\sigma] = 2\text{erf}(k)$ and $P[|X - \mu_X| \geq k\sigma] = 1 - 2\text{erf}(k)$, where $\text{erf}(k)$ is defined in Equation 2.4-12. Using Table 2.4-1 and Equations 4.4-3 and 4.4-4, we obtain Table 4.4-1.

From Table 4.4-1 we see that the Chebyshev bound is not very good; however, it must be recalled that the bound applies to *any* RV X as long as σ^2 exists.

There are a number of extensions of the Chebyshev inequality‡. We consider such an extension in what follows.

Markov Inequality

Consider an RV X for which $f_X(x) = 0$ for $x < 0$. Then X is called a nonnegative RV and the Markov inequality applies:

$$P[X \geq \delta] \leq \frac{E[X]}{\delta}. \tag{4.4-5}$$

In contrast to the Chebyshev bound, which involves both the mean and variance this bound involves only the mean of X.

†The Chebyshev inequality is not very useful when k or δ is small.

‡See Davenport [4-2, p. 256]

Table 4.4-1

k	$P[\lvert X-\overline{X}\rvert < k\sigma]$	CB	$P[\lvert X-\overline{X}\rvert > k\sigma]$	CB
0	0	0	1	1
0.5	0.383	0	0.617	1
1.0	0.683	0	0.317	1
1.5	0.866	0.556	0.134	0.444
2.0	0.955	0.750	0.045	0.250
2.5	0.988	0.840	0.012	0.160
3.0	0.997	0.889	0.003	0.111

Proof of Equation 4.4-5

$$E[X] = \int_0^\infty x f_X(x)\,dx \geq \int_\delta^\infty x f_X(x)\,dx \geq \delta \int_\delta^\infty f_X(x)\,dx$$
$$\geq \delta P[X \geq \delta]$$

whence Equation 4.4-5 follows. Equation 4.4-5 puts a bound on what fraction of a population can exceed δ. ■

Example 4.4-2

(*bad resistors*) Assume that in the manufacturing of very low grade electrical 1000-ohm resistors the average resistance, as determined by a statistical analysis of measurements, is indeed 1000 ohms but there is a large variation about this value. If all resistors over 1500 ohms are to be discarded, what is the maximum fraction of resistors to meet such a fate?

Solution With $\mu_X = 1000$, and $\delta = 1500$, we obtain

$$P[X \geq 1500] \leq \frac{1000}{1500} = 0.67.$$

Thus, if nothing else, the manufacturer has the assurance that the percentage of discarded resistors cannot exceed 67 percent of the total.

The Schwarz Inequality

We have already encountered the probabilistic form of the Schwarz† inequality in Equation 4.3-17 repeated here as

$$\text{Cov}^2[X,Y] \leq E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]$$

with equality if and only if Y is a linear function of X. Upon taking the square root of both sides of this inequality, we have that the magnitude of covariance between two RVs is always upper bounded by the square root of the product of the two variances

$$|\text{Cov}[X,Y]| \leq \left(\sigma_X^2 \sigma_Y^2\right)^{1/2}.$$

†H. Amandus Schwarz (1843–1921), German mathematician.

In later work we shall need another version of the Schwarz inequality that is commonly used in obtaining results in signal processing and stochastic processes. Consider two nonrandom (deterministic) functions h and g not necessarily real valued. Define the *norm* of an ordinary function f by

$$\|f\| \triangleq \left(\int_{-\infty}^{\infty} |f(x)|^2\,dx\right)^{1/2}, \tag{4.4-6}$$

whenever the integral exists. Also define the *scalar* or *inner* product of h with g, denoted by (h,g), as

$$\begin{aligned}(h,g) &\triangleq \int_{-\infty}^{\infty} h(x)g^*(x)\,dx \\ &= (g,h)^*.\end{aligned} \tag{4.4-7}$$

The deterministic form of the Schwarz inequality is then

$$|(h,g)| \le \|h\|\|g\| \tag{4.4-8}$$

with equality if and only if h is proportional to g, that is, $h(x) = ag(x)$ for some constant a. The proof of Equation 4.4-8 is obtained by considering the norm of $\lambda h(x)+g(x)$ as a function of the variable λ,

$$\|\lambda h(x) + g(x)\|^2 = |\lambda|^2\|h\|^2 + \lambda(h,g) + \lambda^*(h,g)^* + \|g\|^2 \ge 0. \tag{4.4-9}$$

If we let

$$\lambda = -\frac{(h,g)^*}{\|h\|^2} \tag{4.4-10}$$

Equation 4.4-8 follows. In the special case where h and g are real functions of real RVs, that is, $h(X)$, $g(X)$, Equation 4.4-8 still is valid provided that the definitions of norm and inner product are modified as follows:

$$\|h\|^2 \triangleq \int_{-\infty}^{\infty} h^2(x)f_X(x)\,dx = E[h^2(X)] \tag{4.4-11}$$

$$(h,g) \triangleq \int_{-\infty}^{\infty} h(x)g(x)f_X(x)\,dx = E[h(X)g(X)] \tag{4.4-12}$$

whence we obtain

$$|E[h(X)g(X)]| \le (E[h^2(X)])^{1/2}(E[g^2(X)])^{1/2}. \tag{4.4-13}$$

Law of large numbers. A very important application of Chebyshev's inequality is to prove the so-called weak *Law of Large Numbers* (LLN) that gives conditions under which a sample mean converges to the ensemble mean.

Example 4.4-3

(weak *law of large numbers*) Let $X_1, \ldots, X_n$ be i.i.d. RVs with mean μ_X and variance σ_X^2. Assume that we don't know the value of μ_X (or σ_X) and thus consider the *sample mean estimator*[†]

$$\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$$

as an estimator for μ_X. We can use the Chebyshev inequality to show that $\hat{\mu}_n$ is asymptotically a perfect estimator for of μ_X. First we compute

$$\begin{aligned} E[\hat{\mu}_n] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n} n\mu_X \\ &= \mu_X. \end{aligned}$$

Next we compute

$$\begin{aligned} \text{Var}[\hat{\mu}_n] &= \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^{n} X_i\right] \\ &= \left(\frac{1}{n^2}\right) n\sigma_X^2 \\ &= \frac{1}{n}\sigma_X^2. \end{aligned}$$

Thus, by the Chebyshev inequality (Equation 4.4-1) we have

$$P[|\hat{\mu}_n - \mu_X| \geq \delta|] \leq \sigma_X^2/n\delta^2.$$

Clearly for any fixed $\delta > 0$, the right side can be made arbitrarily small by choosing n large enough. Thus,

$$\lim_{n\to\infty} P[|\hat{\mu}_n - \mu_X| \geq \delta] = 0$$

for every $\delta > 0$. Note though that for δ small, we may need n quite large to guarantee that the probability of the event $\{|\hat{\mu}_n - \mu_X| \geq \delta\}$ is sufficiently small. This type of convergence is called *convergence in probability* and is treated more extensively in Chapter 8.

The LLN is the theoretical basis for estimating μ_X from measurements. When an experimenter takes the sample mean of n measurements, he is relying on the LLN in order to use the sample mean as an estimate of the unknown mathematical expectation (ensemble average) $E[X] = \mu_X$.

[†]An estimator is a function of the observations $X_1, X_2, \ldots, X_n$ that estimates a parameter of the distribution. Estimators are random variables. When an estimator takes on a particular value, that is, a realization, that number is sometimes called the *estimate*. Estimators are discussed in Chapter 6.

Sometimes inequalities can be derived from properties of the pdf. We illustrate with the following example due to Yongyi Yang.

Example 4.4-4

(*symmetric RVs*) Let the pdf of the real RV X satisfy $f_X(x) = f_X(-x)$; that is, X is symmetrically distributed around zero. Show that $\sigma_X \geq E[|X|]$ with equality if $\text{Var}(|X|) = 0$.

Solution Let $Y \triangleq |X|$. Then $E[Y^2] = E[X^2] = \mu_X^2 + \sigma_X^2 = \sigma_X^2$ since $\mu_X = 0$. Also $E[Y^2] = \mu_Y^2 + \sigma_Y^2 = E^2[|X|] + \sigma_Y^2 = \sigma_X^2$. But $\sigma_Y^2 \geq 0$. Hence $E^2[|X|] \leq \sigma_X^2$ with equality if $\sigma_Y^2 = 0$. Such a case arises when the pdf of X has the form $f_X(x) = \frac{1}{2}[\delta(x-a) + \delta(x+a)]$, where a is some positive number. Then $Y = a$, $\sigma_Y = 0$, and $E[|X|] = \sigma_X$.

Another inequality is furnished by the *Chernoff bound.* We discuss this bound in Section 4.6 after introducing the moment-generating function $M(t)$ in the next section.

4.5 MOMENT-GENERATING FUNCTIONS

The *moment-generating function* (MGF), if it exists, of an RV X is defined by†

$$M(t) \triangleq E[e^{tX}] \tag{4.5-1}$$

$$= \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx, \tag{4.5-2}$$

where t is a complex variable.

For discrete RVs, we can define $M(t)$ using the PMF as

$$M(t) = \sum_i e^{tx_i} P_X(x_i). \tag{4.5-3}$$

From Equation 4.5-2 we see that except for a sign reversal in the exponent, the MGF is the two-sided Laplace transform of the pdf for which there is a known inversion formula. Thus, in general, knowing $M(t)$ is equivalent to knowing $f_X(x)$ and vice versa.

The main reasons for introducing $M(t)$ are (1) it enables a convenient computation of the moments of X; (2) it can be used to estimate $f_X(x)$ from experimental measurements of the moments; (3) it can be used to solve problems involving the computation of the sums of RVs; and (4) it is an important analytical instrument that can be used to demonstrate basic results such as the *Central Limit Theorem.*‡

---

†The terminology varies (see Feller [4-1], p. 411).

‡To be discussed in Section 4.7.

Proceeding formally, if we expand e^{tX} and take expectations, then

$$E[e^{tX}] = E\left[1 + tX + \frac{(tX)^2}{2!} + \ldots + \frac{(tX)^n}{n!} + \ldots\right]$$
$$= 1 + tm_1 + \frac{t^2}{2!}m_2 + \ldots + \frac{t^n}{n!}m_n + \ldots. \tag{4.5-4}$$

Since the moments m_i may not exist, for example, none of the moments above the first exist for the Cauchy pdf, $M(t)$ may not exist. However, if $M(t)$ does exist, computing any moment is easily obtained by differentiation. Indeed, if we allow the notation

$$M^{(k)}(0) \triangleq \frac{d^k}{dt^k}(M(t))\bigg|_{t=0},$$

then

$$m_k = M^{(k)}(0) \qquad k = 0, 1, \ldots. \tag{4.5-5}$$

Example 4.5-1

(*Gaussian MGF*) Let $X : N(\mu, \sigma^2)$. Its MGF is then given as

$$M_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) e^{tx}\,dx. \tag{4.5-6}$$

Using the procedure known as "completing the square"† in the exponent, we can write Equation 4.5-6 as

$$M_X(t) = \exp[\mu t + \sigma^2 t^2/2]$$
$$\times \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x - (\mu + \sigma^2 t))^2\right) dx.$$

But the factor on the second line is unity since it is the integral of a Gaussian pdf with mean $\mu + \sigma^2 t$ and variance σ^2. Hence the Gaussian MGF is

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2), \tag{4.5-7}$$

from which we obtain

$$M_X^{(1)}(0) = \mu$$
$$M_X^{(2)}(0) = \mu^2 + \sigma^2.$$

†See "Completing the square" in Appendix A.

Example 4.5-2

(*MGF of binomial*) Let B be a binomial RV with parameters n (number of tries), p (probability of a success per trial), and $q = 1 - p$. Then the MGF is given as

$$\begin{aligned} M_B(t) &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} [e^t p]^k q^{n-k} \\ &= (pe^t + q)^n. \end{aligned} \tag{4.5-8}$$

We obtain

$$\begin{aligned} M_B^{(1)}(0) &= np = \mu \\ M_B^{(2)}(0) &= \{npe^t(pe^t + q)^{n-1} + n(n-1)p^2 e^{2t}(pe^t + q)^{n-2}\}_{t=0} \\ &= npq + \mu^2. \end{aligned} \tag{4.5-9}$$

Hence

$$\mathrm{Var}[B] = npq. \tag{4.5-10}$$

Example 4.5-3

(*MGF of geometric distribution*) Let X follow the geometric distribution. Then the PMF is $P_X(n) = a^n(1-a)u(n), n = 0, 1, 2, \ldots$ and $0 < a < 1$. The MGF is computed as

$$\begin{aligned} M_X(t) &= \sum_{n=0}^{\infty} (1-a) a^n e^{tn} \\ &= (1-a) \sum_{n=0}^{\infty} (ae^t)^n = \frac{1-a}{1-ae^t}. \end{aligned}$$

Then the mean μ is computed from $\mu = M_X'(0) = (1-a)(1-ae^t)^{-2} ae^{-t}|_{t=0} = a/(1-a)$.

We make the observation that if all the moments exist and are known, then $M(t)$ is known as well (see Equations 4.5-4 and 4.5-2). Since $M_X(t)$ is related to $f_X(x)$ through the Laplace transform, we can, in principle at least, determine $f_X(x)$ from its moments if they exist.† In practice, if X is the RV whose pdf is desired and X_i represents our ith observation of X, then we can *estimate* the rth moment of X, m_r, from

$$\widehat{m}_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r, \tag{4.5-11}$$

†For some distributions not all moments exist. For example, as stated earlier for the Cauchy distribution, all moments above the first do not exist.

where $\widehat{m}_r$ is called the *r-moment estimator* and is an RV, and n is the number of observations. Even though $\widehat{m}_r$ is an RV, its variance becomes small as n becomes large. So for n large enough, we can have confidence that $\widehat{m}_r$ is reasonably close to m_r (a deterministic quantity, that is, not an RV).

The joint MGF $M_{XY}(t_1, t_2)$ of two RVs X and Y is defined by

$$\begin{aligned} M_{XY}(t_1,t_2) &\triangleq E[e^{(t_1X+t_2Y)}] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(t_1x+t_2y)f_{XY}(x,y)\,dx\,dy. \end{aligned} \tag{4.5-12}$$

Proceeding as we did in Equation 4.5-4, we can establish with the help of a power series expansion that

$$M_{XY}(t_1,t_2) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{t_1^i t_2^j}{i!j!}m_{ij}, \tag{4.5-13}$$

where m_{ij} is defined in Equation 4.3-13. Using the notation

$$M_{XY}^{(l,n)}(0,0) \triangleq \left.\frac{\partial^{l+n}M_{XY}(t_1,t_2)}{\partial t_1^l \partial t_2^n}\right|_{t_1=t_2=0,}$$

we can show from Equation 4.5-12 or 4.5-13 that

$$m_{ln} = M_{XY}^{(l,n)}(0,0). \tag{4.5-14}$$

In particular

$$M_{XY}^{(1,0)}(0,0) = \mu_X, \qquad M_{XY}^{(0,1)}(0,0) = \mu_Y \tag{4.5-15}$$

$$M_{XY}^{(2,0)}(0,0) = E[X^2], \qquad M_{XY}^{(0,2)}(0,0) = E[Y^2] \tag{4.5-16}$$

$$M_{XY}^{(1,1)}(0,0) = m_{11} = \text{Cov}[X,Y] + \mu_X\,\mu_Y. \tag{4.5-17}$$

4.6 CHERNOFF BOUND

The Chernoff bound furnishes an upper bound on the tail probability $P[X \geq a]$, where a is some prescribed constant. First note that $u(x-a) \leq e^{t(x-a)}$ for any $t > 0$. Assume that X is a continuous RV. Then

$$\begin{aligned} P[X \geq a] &= \int_a^{\infty} f_X(x)\,dx \\ &= \int_{-\infty}^{\infty} f_X(x)u(x-a)\,dx \end{aligned} \tag{4.6-1}$$

and, by the observation made above, it follows that

$$P[X \geq a] \leq \int_{-\infty}^{\infty} f_X(x)e^{t(x-a)}\,dx \tag{4.6-2}$$

and this must hold for any $t > 0$. But, from Equation 4.5-2,

$$\int_{-\infty}^{\infty} f_X(x)e^{t(x-a)}\,dx = e^{-at}M_X(t), \tag{4.6-3}$$

where the subscript has been added to emphasize that the MGF is associated with X. Combining Equations 4.6-3 and 4.6-2, we obtain

$$P[X \geq a] \leq e^{-at}M_X(t). \tag{4.6-4}$$

The tightest bound, which occurs when the right-hand side is minimized with respect to t, is called the Chernoff bound. We illustrate with some examples.

Example 4.6-1

(*Chernoff bound to Gaussian*) Let $X : N(\mu, \sigma^2)$ and consider the Chernoff bound on $P[X \geq a]$, where $a > \mu$. From Equations 4.5-7 and 4.6-3 we obtain

$$P[X \geq a] \leq e^{-(a-\mu)t+\sigma^2 t^2/2}.$$

The minimum of the right-hand side is obtained by differentiating with respect to t and occurs when $t = (a-\mu)/\sigma^2$. Hence the Chernoff bound is

$$P[X \geq a] \leq e^{-(a-\mu)^2/2\sigma^2}. \tag{4.6-5}$$

The Chernoff bound can be derived for discrete RVs also. For example, assume that an RV X takes values $X = i$, $i = 0, 1, 2, \ldots$, with probabilities $P[X = i] \triangleq P_X(i)$. For any integers n, k, define

$$u(n-k) = \begin{cases} 1, & n \geq k, \\ 0, & \text{otherwise.} \end{cases}$$

If follows, therefore, that

$$\begin{aligned} P[X \geq k] &= \sum_{n=k}^{\infty} P_X(n) \\ &= \sum_{n=0}^{\infty} P_X(n)u(n-k) \\ &\leq \sum_{n=0}^{\infty} P_X(n)e^{t(n-k)} \qquad \text{for } t \geq 0. \end{aligned}$$

The last line follows from the fact that

$$e^{t(n-k)} \geq u(n-k) \qquad \text{for } t \geq 0.$$

We note that

$$\sum_{n=0}^{\infty} P_X(n)e^{t(n-k)} = e^{-tk}\sum_{n=0}^{\infty} P_X(n)e^{tn}$$

$$= e^{-tk}M_X(t) \qquad \text{(by Equation 4.5-3)}.$$

Hence we establish the result

$$P[X \geq k] \leq e^{-tk}M_X(t). \tag{4.6-6}$$

As before, the Chernoff bound is determined by minimizing the right-hand side of Equation 4.6-6. We illustrate with an example.

Example 4.6-2

(*Chernoff bound for Poisson*) Let X be a Poisson RV with parameter $a > 0$. Compute the Chernoff bound for $P_X(k)$, where $k > a$. From homework problem 4.39 we find the MGF

$$M_X(t) = e^{a[e^t-1]}$$

and

$$e^{-tk}M_X(t) = e^{-a}e^{[ae^t-kt]}.$$

By setting

$$\frac{d}{dt}[e^{-tk}M_X(t)] = 0,$$

we find that the minimum is reached when $t = t_m$, where

$$t_m = \ln\frac{k}{a}.$$

Thus with $a = 2$ and $k = 5$, we find

$$P[X \geq 5] \leq e^{-2}\exp[5 - 5\ln(5/2)]$$
$$\leq 0.2.$$

4.7 CHARACTERISTIC FUNCTIONS

If in Equation 4.5-1 we replace the parameter t by $j\omega$, where $j \triangleq \sqrt{-1}$, we obtain the *characteristic function* (CF) of X defined by

$$\Phi_X(\omega) \triangleq E[e^{j\omega X}]$$

$$= \int_{-\infty}^{\infty} f_X(x)e^{j\omega x}\,dx, \tag{4.7-1}$$

which, except for a minus sign difference in the exponent, we recognize as the Fourier transform of $f_X(x)$. For discrete RVs we can define $\Phi_X(\omega)$ in terms of the PMF by

$$\Phi_X(\omega) = \sum_i e^{j\omega x_i} P_X(x_i). \tag{4.7-2}$$

For our purposes, the CF has all the properties of the MGF. The Fourier transform is widely used in statistical communication theory, and since the inversion of Equation 4.7-1 is often easy to achieve, either by direct integration or through the availability of extensive tables of Fourier transforms (e.g., [4-7]), the CF is widely used to solve problems involving the computation of the sums of independent RVs. We have seen that the pdf of the sum of independent RVs involves the convolution of their pdf's. Thus if $Z = X_1 + \ldots + X_N$, where $X_i, i = 1, \ldots, N$, are independent RVs, the pdf of Z is furnished by

$$f_Z(z) = f_{X_1}(z) * f_{X_2}(z) * \ldots * f_{X_N}(z), \tag{4.7-3}$$

that is, the repeated convolution product.

The actual evaluation of Equation 4.7-3 can be tedious. However, we know from our studies of Fourier transforms that the Fourier transform of a convolution product is the product of the individual transforms. We illustrate the use of CFs in the following examples.

Example 4.7-1

(*CF of sum*) Let $Z \triangleq X_1 + X_2$ with $f_{X_1}(x)$, $f_{X_2}(x)$, and $f_Z(z)$ denoting the pdf's of X_1, X_2, and Z, respectively. Show that $\Phi_Z(\omega) = \Phi_{X_1}(\omega)\Phi_{X_2}(\omega)$.

Solution From the main result of Section 3.3 (Equation 3.3-15), we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(z - x)\, dx$$

and the corresponding CF

$$\begin{aligned}\Phi_Z(\omega) &= \int_{-\infty}^{\infty} e^{j\omega z} \left[\int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(z - x)\, dx\right] dz \\ &= \int_{-\infty}^{\infty} f_{X_1}(x) \int_{-\infty}^{\infty} f_{X_2}(z - x) e^{j\omega z} dx\, dz.\end{aligned}$$

With a change of variable $\alpha \triangleq z - x$, we obtain the CF of the sum Z as

$$\Phi_Z(\omega) = \Phi_{X_1}(\omega)\Phi_{X_2}(\omega).$$

This result can be extended to N RVs by induction. Thus if $Z = X_1 + \cdots + X_N$, then the CF of Z would be

$$\Phi_Z = \Phi_{X_1}(\omega)\Phi_{X_2}(\omega) \ldots \Phi_{X_N}(\omega).$$

Example 4.7-2

(*CF of i.i.d. sum*) Let $X_i, i = 1, \ldots, N$, be a sequence of i.i.d. RVs with $X: N(0,1)$. Compute the pdf of

$$Z \triangleq \sum_{i=1}^{N} X_i.$$

Solution The pdf of Z can be computed by Equation 4.7-3. On the other hand, with $\Phi_{X_i}(\omega)$ denoting the CF of X_i, we have

$$\Phi_Z(\omega) = \Phi_{X_1}(\omega) \times \ldots \times \Phi_{X_N}(\omega). \tag{4.7-4}$$

However, since the X_i's are i.i.d. $N(0,1)$, the CFs of all the X_is are the same, and we define $\Phi_X(\omega) \triangleq \Phi_{X_1}(\omega) = \ldots = \Phi_{X_N}(\omega)$. Thus,

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} e^{j\omega x}\, dx. \tag{4.7-5}$$

By completing the squares in the exponent, we obtain

$$\begin{aligned}\Phi_X(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[x^2 - 2j\omega x + (j\omega)^2 - (j\omega)^2]}\, dx \\ &= e^{-\frac{\omega^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x - j\omega)^2}\, dx.\end{aligned}$$

But the integral can be regarded as the area under a "Gaussian pdf" with "mean" $j\omega$ and hence its value is unity†. Thus we obtain the CF of X as

$$\Phi_X(\omega) = e^{-\frac{\omega^2}{2}}$$

and so the CF of Z is

$$\Phi_Z(\omega) = [\Phi_X(\omega)]^n = e^{-\frac{1}{2}n\omega^2}. \tag{4.7-6}$$

From the form of $\Phi_Z(\omega)$ we deduce that $f_Z(z)$ must also be Gaussian. To obtain $f_Z(z)$ we use the Fourier inversion formula:

$$f_Z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_Z(\omega) e^{-j\omega z}\, d\omega. \tag{4.7-7}$$

Inserting Equation 4.7-6 into Equation 4.7-7 and manipulating terms enables us to obtain

$$f_Z(z) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{1}{2}(z^2/n)}.$$

Hence $f_Z(z)$ is indeed Gaussian. The variance of Z is n, and its mean is zero.

†While this result is not obvious, it can be be rigorously demonstrated using integration in the complex plane.

Example 4.7-3

(*CF of sum of uniform RVs*) Consider two independent RVs X and Y with common pdf

$$f_X(x) = f_Y(x) = \frac{1}{a}\text{rect}\left(\frac{x}{a}\right).$$

Compute the pdf of $Z \triangleq X + Y$ using CFs.

Solution We can, of course, compute $f_Z(z)$ by convolving $f_X(x)$ and $f_Y(y)$. However, using CFs, we obtain $f_Z(z)$ from

$$f_Z(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_X(\omega)\Phi_Y(\omega)e^{-j\omega z}d\omega,$$

where

$$\Phi_X(\omega)\Phi_Y(\omega) = \Phi_Z(\omega).$$

Since the pdf's of X and Y are the same, we can write

$$\begin{aligned}\Phi(\omega) &\triangleq \Phi_X(\omega) = \Phi_Y(\omega)\\ &= \frac{1}{a}\int_{-a/2}^{a/2} e^{j\omega x}\,dx\\ &= \frac{\sin(a\omega/2)}{a\omega/2}.\end{aligned}$$

Hence

$$\Phi_Z(\omega) = \left(\frac{\sin(a\omega/2)}{a\omega/2}\right)^2 \tag{4.7-8}$$

and

$$\begin{aligned}f_Z(z) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_Z(\omega)e^{-j\omega z}d\omega\\ &= \frac{1}{a}\left(1 - \frac{|z|}{a}\right)\text{rect}\left(\frac{z}{2a}\right),\end{aligned} \tag{4.7-9}$$

which is shown in Figure 4.7-1. The easiest way to obtain the result in Equation 4.7-9 is to look up the Fourier transform (or its inverse) of Equation 4.7-8 in a table of elementary Fourier transforms.

As in the case of MGFs, we can compute the moments from the CFs by differentiation, provided that these exist. If we expand $\exp(j\omega X)$ into a power series and take the expectation, we obtain

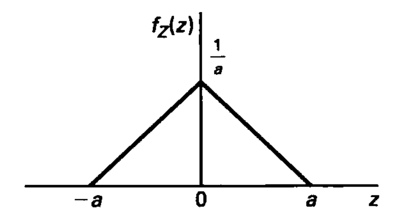


Figure 4.7-1 The pdf of $Z = X + Y$ when X and Y are independently, identically, and uniformly distributed in $(-a/2, a/2)$.

$$\Phi_X(\omega) = E[e^{j\omega X}]$$
$$= \sum_{n=0}^{\infty} \frac{(j\omega)^n}{n!} m_n. \tag{4.7-10}$$

From Equation 4.7-10 it is easily established that

$$m_n = \frac{1}{j^n}\Phi_X^{(n)}(0), \tag{4.7-11}$$

where we have used the notation

$$\Phi_N^{(n)}(0) \triangleq \left.\frac{d^n}{d\omega^n}\Phi_X(\omega)\right|_{\omega=0}.$$

Example 4.7-4

(*moment calculation*) Compute the first few moments of $Y = \sin\Theta$ if $\Theta : U[0, 2\pi]$.

Solution We use the result in Equation 4.1-9; that is, if $Y = g(X)$, then

$$\mu_Y = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx.$$

Hence

$$E[e^{j\omega Y}] = \int_{-\infty}^{\infty} e^{j\omega y} f_Y(y)dy$$
$$= \frac{1}{2\pi}\int_0^{2\pi} e^{j\omega \sin\theta} d\theta$$
$$= J_0(\omega),$$

where $J_0(\omega)$ is the Bessel function of the first kind of order zero. A power series expansion of $J_0(\omega)$ gives

$$J_0(\omega) = 1 - \left(\frac{\omega}{2}\right)^2 + \frac{1}{2!2!}\left(\frac{\omega}{2}\right)^4 - \ldots.$$

Hence all the odd-order moments are zero. From Equation 4.7-11 we compute

$$E[Y^2] = m_2 \triangleq (-1)\Phi_X^{(2)}(0) = \tfrac{1}{2}$$
$$E[Y^4] = m_4 = (+1)\Phi_X^{(4)}(0) = \tfrac{3}{8}.$$

Example 4.7-5

(*sum of independent binomials*) Let X and Y be i.i.d. binomial RVs with parameters n and p, that is,

$$P_X(k) = P_Y(k) = \binom{n}{k} p^k q^{n-k}.$$

Compute the PMF of $Z = X + Y$.

Solution Since X and Y take on nonnegative integer values, so must Z. We can solve this problem by (1) convolution of the pdf's, which involves delta functions; (2) *discrete convolution* of the PMFs; and (3) CFs. The discrete convolution for this case is

$$\begin{aligned} P_Z(k) &= \sum_i P_X(i)P_Y(k-i) \\ &= \sum_i \binom{n}{i} p^i q^{n-i} \binom{n}{k-i} p^{k-i} q^{n-(k-i)} \\ &= p^k q^{2n-k} \sum_i \binom{n}{i}\binom{n}{k-i}, \quad \text{for } k = 0, 1, \ldots, 2n. \end{aligned}$$

The trouble is that we may not immediately recognize the closed form of the sum of products of binomial coefficients.† The computation of the PMF of Z by CFs is very simple. First observe that

$$\begin{aligned} \Phi_X(\omega) = \Phi_Y(\omega) &= \sum_{k=0}^{n} e^{j\omega k} \binom{n}{k} p^k q^{n-k} \\ &= (pe^{j\omega} + q)^n. \end{aligned}$$

Thus, by virtue of the independence of X and Y, we obtain the CF

$$\begin{aligned} \Phi_Z(\omega) &= E[\exp j\omega(X+Y)] \\ &= E[\exp(j\omega X)]E[\exp(j\omega Y)] \\ &= (pe^{j\omega} + q)^{2n}. \end{aligned}$$

Thus Z is binomial with parameters $2n$ and p, that is,

$$P_Z(k) = \binom{2n}{k} p^k q^{2n-k}, \quad \text{for } k = 0, \ldots, 2n.$$

†Recall that we ran into this problem in Example 3.3-9 in Chapter 3.

As a by-product of the computation of $P_Z(k)$ by CFs, we obtain the result that

$$\binom{2n}{k} = \sum_{i=0}^{n} \binom{n}{i} \binom{n}{k-i}.$$

An extension of this result is the following: If $X_1, X_2, \ldots, X_N$ are i.i.d. binomials with parameters n, p, then $Z = \sum_{i=1}^{N} X_i$ is binomial with parameters Nn, p. Regardless of how large N gets, *Z remains a discrete RV with a binomial PMF.*†

Example 4.7-6

(*variance of Poisson*) Here we calculate the CF of a Poisson RV and use it to determine the variance. Let the RV K be Poisson distributed with PMF

$$P_K(k) = \frac{\alpha^k}{k!} e^{-\alpha} u(k), \quad \alpha > 0.$$

Then the CF is given as

$$\begin{aligned} \Phi_K(\omega) &= \sum_{k=0}^{\infty} \frac{\alpha^k}{k!} e^{-\alpha} e^{j\omega k} \\ &= \sum_{k=0}^{\infty} \frac{\left(\alpha e^{j\omega}\right)^k}{k!} e^{-\alpha} \\ &= \exp\left[\alpha\left(e^{j\omega} - 1\right)\right]. \end{aligned}$$

Now $m_2 = E[K^2] = \frac{1}{j^2}\Phi_K^{(2)}(0) = -\Phi_K^{(2)}(0)$. Taking the indicated derivatives, we get

$$\Phi_K^{(1)}(\omega) = \Phi_K(\omega)\alpha j e^{+j\omega}$$

and

$$\begin{aligned} \Phi_K^{(2)}(\omega) &= \Phi_K(\omega)\alpha j^2 e^{+j\omega} + \Phi_K^{(1)}(\omega)\alpha j e^{+j\omega} \\ &= -\Phi_K(\omega)\alpha e^{+j\omega} + \Phi_K(\omega)\left(\alpha j e^{+j\omega}\right)^2. \end{aligned}$$

So $\Phi_K^{(2)}(0) = -1 \times \alpha - 1 \times \alpha^2$. Hence $\mu_2 = \alpha + \alpha^2$. Then since the mean is $\mu = a$, the variance must be

$$\begin{aligned} \sigma^2 &= m_2 - \mu^2 \\ &= \alpha + \alpha^2 - \alpha^2 \\ &= \alpha. \end{aligned}$$

The variance of the Poisson RV thus equals its mean value.

---

†Recall this statement for future reference in connection with the *Central Limit Theorem.*

Note that since the variance of the Poisson RV equals its mean, the standard deviation is the square root of the mean. Therefore for large mean values, the distribution becomes relatively concentrated around the mean. Another point is that unlike the Normal distribution the mean and variance of the Poisson RV are not independent parameters i,e., they cannot be freely chosen.

Example 4.7-7

(*a fair game?*) A lottery game called "three players for six hits" is played as follows. A bettor bets the bank that three baseball players of the bettor's choosing will get a combined total of six hits or more in the games in which they play. Many combinations can lead to a win; for example, player A can go hitless in his game, but player B can collect three hits in his game, and player C can collect three hits in his game. The players can be on the same team or on different teams. The bet is at even odds and the bettor receives back $2 on a bet of $1 in case of a win. Is this a "fair" game, that is, is the probability of a win close to one-half?

Solution Let X_1, X_2, X_3 denote the number of hits by players A, B, C, respectively. Clearly X_1, X_2, X_3 are individually binomial. The total number of hits is $Y = \sum_{i=1}^{3} X_i$. We wish to compute $P[Y \geq 6]$. To simplify the problem, assume that each player bats five times per game, and their batting averages are the same, say 300 (for those unfamiliar with baseball nomenclature, this means that the probability of getting a hit while batting is 0.3). Then from the results of Example 4.7-5, we find Y is binomial with parameters $n = 15$, $p = 0.3$. Thus,

$$\begin{aligned} P[Y \geq 6] &= \sum_{i=6}^{i=15} \binom{15}{i} (0.3)^i (0.7)^{15-i} \\ &\approx \text{erf}(6.76) - \text{erf}(0.56) \\ &\approx 0.29. \end{aligned}$$

In arriving at this result, we used the Normal approximation to the binomial as suggested in Chapter 1, Section 1.11. The bettor has less than a *one-third chance* of winning. Despite the poor odds, the game can be modified to be fairer to the bettor. Define the RV G as the *gain* to the bettor and define a fair game as one in which the expected gain is zero. Then if the bettor were to receive winnings of \$2.45 per play instead of \$1, we would find that $E[G] = \$2.45 \times 0.29 - \$1 \times 0.71 \approx 0$. Of course if $E[G] > 0$, then in a sense, the game favors the bettor. Some people play the state lottery using this criterion.

Joint Characteristic Functions

As in the case of joint MGFs we can define the joint CF by

$$\Phi_{X_1 \ldots X_N}(\omega_1, \omega_2, \ldots, \omega_N) = E\left[\exp\left(j \sum_{i=1}^{N} \omega_i X_i\right)\right]. \tag{4.7-12}$$

By the Fourier inversion property, the joint pdf is the inverse Fourier transform (with a sign reversal) of $\Phi_{X_1 \ldots X_N}(\omega_1, \ldots, \omega_N)$. Thus,

$$f_{X_1 \dots X_N}(x_1, \dots, x_N) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \Phi_{X_1 \dots X_N}(\omega_1, \dots, \omega_N)$$
$$\times \exp\left(-j \sum_{i=1}^{N} \omega_i x_i\right) d\omega_1\, d\omega_2 \dots d\omega_N. \tag{4.7-13}$$

We can obtain the moments by differentiation. For instance, with X, Y denoting any two RVs ($N = 2$) we have

$$m_{rk} \triangleq E[X^r Y^k] = (-j)^{r+k} \Phi_{XY}^{(r,k)}(0,0), \tag{4.7-14}$$

where

$$\Phi_{XY}^{(rk)}(0,0) \triangleq \left. \frac{\partial^{r+k} \Phi_{XY}(\omega_1, \omega_2)}{\partial \omega_1^r \partial \omega_2^k} \right|_{\omega_1 = \omega_2 = 0}. \tag{4.7-15}$$

Finally for discrete RVs we can define the joint CF in terms of the joint PMF. For instance for two RVs, X and Y, we obtain

$$\Phi_{XY}(\omega_1, \omega_2) \triangleq \sum_i \sum_j e^{j(\omega_1 x_i + \omega_2 y_j)} P_{XY}(x_i, y_j). \tag{4.7-16}$$

Example 4.7-8

(*joint CF of i.i.d. Normal RVs*) Compute the joint characteristic function of X and Y if

$$f_{XY} = \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x^2 + y^2)\right].$$

Solution Applying the definition in Equation 4.7-12, we get

$$\Phi_{XY}(\omega_1, \omega_2) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} e^{j\omega_1 x + j\omega_2 y}\, dx\, dy.$$

Completing the squares in both x *and* y, we get

$$\Phi_{XY}(\omega_1, \omega_2) = e^{-\frac{1}{2}(\omega_1^2 + \omega_2^2)} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[x^2 - 2j\omega_1 x + (j\omega_1)^2]} \frac{dx}{\sqrt{2\pi}}$$
$$\times \int_{-\infty}^{\infty} e^{-\frac{1}{2}[y^2 - 2j\omega_2 y + (j\omega_2)^2]} \frac{dy}{\sqrt{2\pi}}$$
$$= e^{-\frac{1}{2}(\omega_1^2 + \omega_2^2)} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x - j\omega_1)^2} \frac{dx}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(y - j\omega_2)^2} \frac{dy}{\sqrt{2\pi}}$$
$$= e^{-\frac{1}{2}(\omega_1^2 + \omega_2^2)},$$

since the integrals are the areas under unit-variance Gaussian curves.

Example 4.7-9

(*joint CF of two discrete RVs*) Compute the joint CF of the discrete RVs X and Y if the joint PMF is

$$P_{XY}(k,l) = \begin{cases} \frac{1}{3}, & k = l = 0, \\ \frac{1}{6}, & k = \pm 1, l = 0, \\ \frac{1}{6}, & k = l = \pm 1, \\ 0, & \text{else.} \end{cases}$$

Solution Using Equation 4.7-16 we obtain

$$\begin{aligned} \Phi_{XY}(\omega_1, \omega_2) &= \sum_{k=-1}^{1} \sum_{l=-1}^{1} e^{j(\omega_1 k + \omega_2 l)} P_{XY}(k,l) \\ &= \frac{1}{3} + \frac{1}{3}\cos\omega_1 + \frac{1}{3}\cos(\omega_1 + \omega_2). \end{aligned}$$

From Equations 4.7-14 and 4.7-15 we obtain, since $\mu_X = \mu_Y = 0$,

$$\begin{aligned} \sigma_X^2 \triangleq m_{20} &= -(-j)^2[\cos\omega_1 + \cos(\omega_1 + \omega_2)]\, \frac{1}{3}\Big|_{\omega_1=\omega_2=0} \\ &= \frac{2}{3}; \\ \sigma_Y^2 \triangleq m_{02} &= -(-j)^2\, \frac{1}{3}\cos(\omega_1 + \omega_2)\Big|_{\omega_1=\omega_2=0} \\ &= \frac{1}{3}; \\ m_{11} &= -(-j)^2\, \frac{1}{3}\cos(\omega_1 + \omega_2)\Big|_{\omega_1=\omega_0=0} \\ &= \frac{1}{3}. \end{aligned}$$

Hence the correlation coefficient ρ is computed to be

$$\rho = \frac{m_{11}}{\sigma_X \sigma_Y} = \frac{\frac{1}{3}}{\sqrt{\frac{2}{3}}\sqrt{\frac{1}{3}}} = \frac{1}{\sqrt{2}} = 0.707.$$

Example 4.7-10

(*joint CF of correlated Normal RVs*) As another example we compute the joint CF of X and Y with

$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right).$$

To solve this problem we use two facts:

(1) A zero-mean Gaussian RV Z with variance σ_Z^2 has CF

$$E[e^{j\omega Z}] = \exp\left[-\frac{1}{2}\sigma_Z^2\omega^2\right] \tag{4.7-17}$$

and, in particular, with $\omega = 1$,

$$E[e^{jZ}] = \exp\left[-\frac{1}{2}\sigma_Z^2\right]. \tag{4.7-18}$$

Proof of fact (1) Use the definition of the CF with $f_Z(z) = (2\pi\sigma_Z^2)^{-1/2}\exp\left(-\frac{1}{2}\frac{z^2}{\sigma_Z^2}\right)$ and apply the complete-the-square technique described in Appendix A.

(2) If X and Y are zero-mean jointly Gaussian RVs, then for any real ω_1, ω_2, the RVs

$$\begin{aligned} Z &\triangleq \omega_1 X + \omega_2 Y \\ W &\triangleq X \end{aligned}$$

are jointly Gaussian and, as a direct by-product, the marginal density of Z is Gaussian.

Proof of fact (2) Simply use Equation 3.4-11 or 3.4-12 to compute $f_{ZW}(z, w)$. One easily finds that Z, W are jointly Gaussian and that, therefore, the marginal pdf of Z alone is Gaussian with $\overline{Z} = 0$. The variance of Z is computed as

$$\begin{aligned} \text{Var}(Z) &= E[(\omega_1 X + \omega_2 Y)^2] \\ &= \omega_1^2 \text{Var}[X] + \omega_2^2 \text{Var}[Y] + 2\omega_1\omega_2\overline{XY}. \end{aligned}$$

With $\sigma_X^2 = \sigma_Y^2 \triangleq 1$, we obtain $\sigma_Z^2 = \omega_1^2 + \omega_2^2 + 2\omega_1\omega_2\rho$.

Finally recalling that $Z = \omega_1 X + \omega_2 Y$ and using Equation 4.7-18, we write

$$E[e^{j(\omega_1 X + \omega_2 Y)}] = e^{-\frac{1}{2}(\omega_1^2 + \omega_2^2 + 2\omega_1\omega_2\rho)}. \tag{4.7-19}$$

Equation 4.7-19 is the joint CF of two zero-mean, unity variance correlated Gaussian RVs. When $\rho = 0$, the RVs become uncorrelated and therefore independent and we obtain the result in Example 4.7-8.

The extension to more than two discrete RVs is straightforward, although the notation becomes a little clumsy, unless matrices are introduced.

The Central Limit Theorem

It is sometimes said that the sum of a large number of RVs tends toward the Normal. Under what conditions is this true? The *Central Limit Theorem* deals with this important point.

Basically the Central Limit Theorem† says that the normalized sum of a large number of mutually independent RVs $X_1, \ldots, X_n$ with zero means and finite variances $\sigma_1^2, \ldots, \sigma_n^2$ tends to the Normal CDF provided that the individual variances σ_k^2, $k = 1, \ldots, n$, are small compared to $\sum_{i=1}^{n} \sigma_i^2$. The constraints on the variances are known as the Lindeberg conditions and are discussed in detail by Feller [4-1, p. 262]. We state a general form of the Central Limit Theorem in the following and furnish a proof for a special case.

Theorem 4.7-1 Let $X_1, \ldots, X_n$ be n mutually independent (scalar) RVs with CDFs $F_{X_1}(x_1)$, $F_{X_2}(x_2), \ldots, F_{X_n}(x_n)$, respectively, such that

$$\mu_{X_k} = 0, \qquad \text{Var}[X_k] = \sigma_k^2$$

and let

$$s_n^2 \triangleq \sigma_1^2 + \ldots + \sigma_n^2.$$

If for a given $\varepsilon > 0$ and n sufficiently large the σ_k satisfy

$$\sigma_k < \varepsilon s_n, \qquad k = 1, \ldots, n,$$

then the normalized sum

$$Z_n \triangleq (X_1 + \ldots + X_n)/s_n$$

converges to the standard Normal CDF, denoted by $1/2 + \text{erf}(z)$, that is, $\lim_{n\to\infty} F_{Z_n}(z) = 1/2 + \text{erf}(z)$. This is called convergence in distribution. ■

A discussion of convergence in distribution is given later in this section.

We now prove a special case of the foregoing.

Theorem 4.7-2 Let X_1, $X_2, \ldots, X_n$ be i.i.d. RVs with $\mu_{X_i} = 0$, and $\text{Var}[X_i] = 1$, $i = 1, \ldots, n$. Then

$$Z_n \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

tends to the Normal in the sense that its CF Φ_{Z_n} satisfies

$$\lim_{n\to\infty} \Phi_{Z_n}(\omega) = e^{-\frac{1}{2}\omega^2},$$

which is the CF of the $N(0, 1)$ RV.

Proof Let $W_i \triangleq X_i/\sqrt{n}$. Also, let $\Phi_{X_i}(\omega)$ and $f_{X_i}(x)$ be the CF and pdf of X, respectively. Then

†First proved by Abraham De Moivre in 1733 for the special case of Bernoulli RVs. A more general proof was furnished by J. W. Lindeberg in *Mathematische Zeitschrift*, vol. 15 (1922), pp. 211–225.

$$\begin{aligned}\Phi_{W_i} &\triangleq E[e^{j\omega W_i}] \\ &= E[e^{j(\omega/\sqrt{n})X_i}] \\ &= \Phi_{X_i}\left(\frac{\omega}{\sqrt{n}}\right).\end{aligned}$$

Since $\Phi_{X_i}(\omega)$ and $\Phi_{W_i}(\omega)$ do not depend on i, we write $\Phi_{X_i}(\omega) \triangleq \Phi_X(\omega)$ and $\Phi_{W_i}(\omega) \triangleq \Phi_W(\omega)$. From calculus we know that any function $\Phi(\omega)$ whose derivative exists in a neighborhood about ω_0 can be represented by a Taylor series

$$\Phi(\omega) = \sum_{l=0}^{\infty} \frac{1}{l!}\Phi^{(l)}(\omega_0)(\omega-\omega_0)^l,$$

where $\Phi^{(l)}(\omega_0)$ is the lth derivative of $\Phi(\omega)$ at ω_0. Moreover, if the derivatives are continuous in the interval $[\omega_0, \omega]$, $\Phi(\omega)$ can be expressed as a finite Taylor series plus a remainder $A_L(\omega)$, that is,

$$\Phi(\omega) = \sum_{l=0}^{L-1} \frac{1}{l!}\Phi^{l}(\omega_0)(\omega-\omega_0)^l + A_L(\omega),$$

where

$$A_L(\omega) \triangleq \frac{1}{L!}\Phi^{L}(\xi)(\omega-\omega_0)^L$$

and ξ is some point in the interval $[\omega_0, \omega]$. Let us apply this result to $\Phi_W(\omega)$ with $\omega_0 = 0$. Then

$$\begin{aligned}\Phi_W(\omega) &= \int_{-\infty}^{\infty} e^{j\omega x/\sqrt{n}} f_X(x)\,dx \\ \Phi_W^{(0)}(0) &= 1 \\ \Phi_W^{(1)}(0) &= \int_{-\infty}^{\infty} j\,\frac{x}{\sqrt{n}} e^{j\omega x/\sqrt{n}} f_X(x)\,dx\Big|_{\omega=0} = 0 \\ \Phi_W^{(2)}(0) &= \int_{-\infty}^{\infty} \left(\frac{jx}{\sqrt{n}}\right)^2 e^{j\omega x/\sqrt{n}} f_X(x)\,dx\Big|_{\omega=0} = -\frac{1}{n}.\end{aligned}$$

Hence

$$\Phi_W(\omega) = 1 - \frac{1}{2n}\omega^2 + \frac{R_2'(\omega)}{n\sqrt{n}},$$

where

$$R_2'(\omega) \triangleq -j\omega^3 \int_{-\infty}^{\infty} x^3 e^{j\xi x/\sqrt{n}} f_X(x)\,dx/6.$$

Since $Z_n = \sum_{i=1}^{n} W_i$, we obtain

$$\Phi_{Z_n}(\omega) = [\Phi_W(\omega)]^n,$$

or

$$\ln \Phi_{Z_n}(\omega) = n \ln \Phi_W(\omega).$$

Now recall that for any h such that $|h| < 1$,

$$\ln(1+h) = h - \frac{h^2}{2} + \frac{h^3}{3} - \ldots.$$

For any *fixed* ω, we can choose an n large enough so that (let $R_2' \triangleq R_2'(\omega)$)

$$\left| -\frac{\omega^2}{2n} + \frac{R_2'}{n\sqrt{n}} \right| < 1.$$

Assuming this to have been done, we can write

$$\begin{aligned}
\ln \Phi_{Z_n}(\omega) &= n \ln \left[1 - \frac{\omega^2}{2n} + \frac{R_2'}{n\sqrt{n}} \right] \\
&\simeq n \left[-\frac{\omega^2}{2n} + \frac{R_2'}{n\sqrt{n}} - \frac{1}{2}\left(-\frac{\omega^2}{2n} + \frac{R_2'}{n\sqrt{n}} \right)^2 + \frac{1}{3}\left(-\frac{\omega^2}{2n} + \frac{R_2'}{n\sqrt{n}} \right)^3 + \cdots \right] \\
&= -\frac{\omega^2}{2} + \text{ terms involving factors of } n^{-1/2}, n^{-1}, n^{-3/2}, \ldots.
\end{aligned}$$

Hence

$$\lim_{n\to\infty} [\ln \Phi_{Z_n}(\omega)] = -\frac{\omega^2}{2}$$

or, equivalently,

$$\lim_{n\to\infty} \Phi_{Z_n}(\omega) = e^{-\omega^2/2},$$

which is the CF of the $N(0,1)$ RV. Note that to argue that $\lim_{n\to\infty} f_{Z_n}(z)$ is the normal pdf we should have to argue that

$$\begin{aligned}
\lim_{n\to\infty} \Phi_{Z_n}(\omega) &\triangleq \lim_{n\to\infty} \left(\int_{-\infty}^{\infty} f_{Z_n}(z) e^{j\omega z} dz \right) \\
&\stackrel{?}{=} \int_{-\infty}^{\infty} \left(\lim_{n\to\infty} f_{Z_n}(z) \right) e^{j\omega z} dz.
\end{aligned}$$

However, the operations of limiting and integrating are not always interchangeable. Hence we cannot say that the pdf of Z_n converges to $N(0,1)$. Indeed we already know from Example 4.7-5 that the sum of n i.i.d. binomial RVs is binomial regardless of how large n is; moreover, the binomial PMF or pdf is a discontinuous function while the Gaussian is continuous and no matter how large n is, this fact cannot be altered. However, the integrals of the binomial pdf, for large n, behave like integrals of the Gaussian pdf. This is why the distribution function of Z_n tends to a Gaussian distribution function but not necessarily to a Gaussian pdf.

The astute reader will have noticed that in the prior development we showed the normal convergence of the CF but not as yet the normal convergence of the CDF. To prove the latter true we can use a continuity theorem† which states the following: Consider a sequence of RVs $Z_i, i = 1, \ldots, n$, with CFs and CDFs $\Phi_i(\omega)$ and $F_i(z)$, $i = 1, \ldots, n$, respectively, with $\Phi(\omega) \triangleq \lim_{n\to\infty} \Phi_n(\omega)$ and $\Phi(\omega)$ continuous at $\omega = 0$; then $F(z) = \lim_{n\to\infty} F_n(z)$. ■

Example 4.7-11

(*application of the Central Limit Theorem [CLT]*) Let $X_i, i = 1, \ldots, n$, be a sequence of i.i.d. RVs with $E[X_i] = \mu_X$ and $\text{Var}[X_i] = \sigma_X^2$. Let $Y \triangleq \sum_{i=1}^n X_i$ where n is large. We wish to compute $P[a < Y \leq b]$ using the CLT. With $Z \triangleq (Y - E[Y])/\sigma_Y$, and $\sigma_Y > 0$,

$$P[a < Y \leq b] = P[a' < Z \leq b'],$$

where

$$a' \triangleq \frac{a - E[Y]}{\sigma_Y}$$

$$b' = \frac{b - E[Y]}{\sigma_Y}$$

and

$$\sigma_Y = \sqrt{n}\sigma_X.$$

Note that Z is a zero-mean, unity variance RV involving the sum of a large number (n assumed large) of i.i.d. RVs. Indeed with some minor manipulations we can write Z as

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu_X}{\sigma_X} \right).$$

Hence

$$P[a' < Z \leq b'] \simeq \frac{1}{\sqrt{2\pi}} \int_{a'}^{b'} e^{-\frac{1}{2}z^2} dz.$$

Although the CLT might be more appropriately called the "Normal convergence theorem," the word *central* in Central Limit Theorem is useful as a reminder that CDFs converge to the normal CDF around the center, that is, around the mean. Although all CDFs converge together at $\pm\infty$, it is in fact in the tails that the CLT frequently gives the poorest estimates of the correct probabilities, if these are small. An illustration of this phenomenon is given in Problem 4.59.

In a type of computer-based engineering analysis called *Monte-Carlo simulation*, it is often necessary to have access to random numbers. There are several random number generators available in software that generate numbers that appear random but in fact are not: They are generated using an algorithm that is completely deterministic and therefore

†See Feller [4-1, p. 508].

they can be duplicated by anyone who has a copy of the algorithm. The numbers, called pseudo-random numbers, are often adequate for situations where not too many random numbers are needed. For situations where a very large number of random numbers are needed, for example, modeling atomic processes, it turns out that it is difficult to find an adequate random number generator. Most will eventually display number sequences that repeat, that is, are periodic, are highly correlated, or show other biases. Note that the alternative, that is, using a naturally random process such as the emission of photons from x-ray sources or the liberation of photoelectrons from valence bands in photodetectors, also suffers from a major problem: We cannot be certain what underlying probability law is truly at work. And even if we knew what law was at work, the very act of counting photons or photoelectrons might bias the distribution of random numbers.

In any case, if we assume that for our purposes the uniform random number generators (URNG) commonly available with most PC software packages are adequate in that they create unbiased realizations of a uniform RV X, the next question is how can we convert uniform random numbers, that is, those that are assumed to obey the uniform pdf in (0, 1), to Gaussian random numbers. For this purpose we can use the CLT as follows. Let X_i represent the ith random number generated by the URNG. Then

$$Z = X_1 + \ldots + X_n$$

will be approximately Gaussian for a reasonably large n (say >10). Note that the pdf of Z is the n-repeated convolution of a unit pulse which starts to look like a Gaussian very quickly everywhere except in the tails. The reason there is a problem in the tails is that Z is confined to the range $0 \leq Z \leq n$ while if Z were a true Gaussian RV, then $-\infty < Z < \infty$.

4.8 ADDITIONAL EXAMPLES

Example 4.8-1

Let $X_i, i = 1, ..., n$, be n i.i.d. Bernoulli RVs with individual PMF:

$$P_{X_i}(x) = \begin{cases} p^x(1-p)^{1-x}, x = 0, 1 \\ 0, else. \end{cases}$$

Show that $Z = \sum_{i=1}^{n} X_i$ is binomial with PMF $b(k; n, p) \triangleq \binom{n}{k} p^k q^{n-k}$.

Solution The CF of the Bernoulli RV is computed as $\Phi_{X_i}(\omega) = \sum_{x=0}^{1} e^{j\omega x} p^x (1-p)^{1-x} = pe^{j\omega} + q$, where $q = 1 - p$. From Equation 4.7-4, we obtain that

$$\Phi_Z(\omega) = \prod_{i=1}^{n} (pe^{j\omega} + 1) = (pe^{j\omega} + 1)^n,$$

which, from Example 4.7-5, we recognize as the CF of the binomial RV with PMF as above.

Example 4.8-2

Let Z be a binomial RV with PMF $b(k;n,p) \triangleq \binom{n}{k} p^k q^{n-k}$, where $n \gg 1$, and consider the event $\{a \leq Z \leq b\}$, where a and b are numbers. Use the CLT to compute $P[a \leq Z \leq b]$.

Solution From Example 4.8-1 we know Z can be resolved as a sum of n i.i.d. Bernoulli RVs. Thus we write $Z = X_1 + \cdots X_n$, where $E[Z] = np$ and Var $[Z]$ $=npq \gg pq$ when n is large. The situation is now ripe for applying Theorem 4.7-1, the Central Limit Theorem. The event $\{a \leq Z \leq b\}$ is identical to the event $\{\frac{a-np}{\sqrt{npq}} \leq \frac{Z-np}{\sqrt{npq}} \leq \frac{b-np}{\sqrt{npq}}\}$. With $a' \triangleq \frac{a-np}{\sqrt{npq}}, b' \triangleq \frac{b-np}{\sqrt{npq}}$, and $Z' \triangleq \frac{Z-np}{\sqrt{npq}}$ the event can be rewritten as $\{a' \leq Z' \leq b'\}$, where Z' is a zero-mean, unit-variance RV. Then from Example 4.7-11, which uses a formula based on the CLT, we get

$$P[a \leq Z \leq b] \cong \frac{1}{\sqrt{2\pi}} \int_{a'}^{b'} \exp[-\frac{1}{2}z^2]dz.$$

In terms of the *standard Normal distribution*, $F_{SN}(x)$, defined in Equation 1.11-3, this result can be written as

$$P[a \leq Z \leq b] \approx \quad F_{SN}\left[\frac{b-np}{\sqrt{npq}}\right] - F_{SN}\left[\frac{a-np}{\sqrt{npq}}\right].$$

The correction factor of 0.5 in the limits in Equation 1.11-5 is insignificant when $n \gg 1$.

Example 4.8-3

Let Z be a binomial RV with mean np and standard deviation $\sqrt{npq}$. Use the Normal approximation furnished by the CLT to compute the probability of the following events: $\{np - \sqrt{npq} \leq Z \leq np + \sqrt{npq}\}$, $\{np - 2\sqrt{npq} \leq Z \leq np + 2\sqrt{npq}\}$, $\{np - 3\sqrt{npq} \leq Z \leq np + 3\sqrt{npq}\}$.

Solution With the change of variable $Z' \triangleq \frac{Z-np}{\sqrt{npq}}$, the three events are converted to $\{-1 \leq Z' \leq 1\}, \{-2 \leq Z' \leq 2\}, \{-3 \leq Z' \leq 3\}$. The RV Z' is zero-mean, unit variance and the Normal approximation furnished by the CLT yields:

$$\begin{aligned} P[-1 \leq Z' \leq 1] &= F_{SN}(1) - F_{SN}(-1) \approx 0.683 \\ P[-2 \leq Z' \leq 2] &= F_{SN}(2) - F_{SN}(-2) \approx 0.954 \\ P[-3 \leq Z' \leq 3] &= F_{SN}(3) - F_{SN}(-3) \approx 0.997. \end{aligned}$$

Note that the last-listed event is (almost) certain to occur. In a thousand repetitions it will on the average fail to occur only three times.

Example 4.8-4

Let $X_i, i = 1, \ldots, 100$, be i.i.d. Poisson RVs with PMF $P[k] = e^{-2}\frac{2^k}{k!}, k = 0, 1, 2, \ldots$. Here k is the number of events in a given interval of time. Let $Z = \sum_{i=1}^{100} X_i$. We note that $E[Z] = 200$ and Var[Z]=200. This situation might reflect the summed data packets collected at a receiver from identical multiple channels. Use the CLT to compute the probability of the event $\{190 \leq Z \leq 210\}$.

Solution Since Z is the sum of a large number of i.i.d. RVs and the variance of any of these is much smaller than the variance of the sum, the CLT permits us to use the Normal approximation to compute the probability of this event. Define the RV $Z' = \frac{Z-200}{14.14}$, which is zero-mean and unity variance. Then, in terms of Z', the event becomes $\{-0.707 \leq Z' \leq 0.707\}$. The Normal approximation yields $F_{SN}(0.707) - F_{SN}(-0.707) \approx 0.52$.

SUMMARY

In this chapter we discussed the various averages of one or more random variables (RVs) and the implication of those averages. We began by defining the average or expected value of an RV X and then showed that the expected value of $Y = g(X)$ could be computed directly from the pdf or PMF of X. We briefly discussed the important notion of conditional expectation and showed how the expected value of an RV could be advantageously computed by averaging over its conditional expectation. We then argued that a single summary number such as the average value, μ_X, of X was insufficient for describing the behavior of X. This led to the introduction of moments, that is, the average of powers of X. We illustrated how moments can be used to estimate pdf's by the maximum entropy principle and introduced the concept of *joint moments.* We showed how the covariance of two RVs could be interpreted as a measure of how well we can predict one RV from observing another using a linear predictor model. By giving a counterexample, we demonstrated that uncorrelatedness does not imply independence of two RVs, the latter being a stronger condition. The joint Gaussian pdf for two RVs was discussed, and it was shown that in the Gaussian case, independence and uncorrelatedness are equivalent. We then introduced the reader to some important bounds and inequalities known as the Chebyshev and Schwarz inequalities and the Chernoff bound and illustrated how these are used in problems in probability.

The second half of the chapter dealt mostly with moment generating functions (MGFs) and characteristic functions (CFs) and the Central Limit Theorem (CLT). We showed how the MGF and CF are essentially the Laplace and Fourier transforms, respectively, of the pdf of an RV and how we could compute all the moments, provided that these exist, from either of these functions. Several properties of these important functions were explored. We illustrated how the CF could be used to solve problems involving the computation of the pdf's of the sums of RVs.

We then discussed the CLT, one of the most important results in probability theory, and the basis for the ubiquitous Normal behavior of many random phenomena. The CLT states that under relatively loose mathematical constraints, the cumulative distribution function (CDF) of the sum of independent RVs tends toward the Normal CDF.

We ended the chapter with additional examples of the use and application of the CLT.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

4.1 Compute the average and standard deviation of the following set: 3.50, 5.61, −2.37, 4.94, −6.25, −1.05, −3.75, 5.81, 2.27, 0.54, 6.11, −2.56.

4.2 Compute $E[X]$ when X is a Bernoulli RV, that is,

$$X = \begin{cases} 1, & P_X(1) = p > 0, \\ 0, & P_X(0) = 1 - p > 0. \end{cases}$$

4.3 Let $X = a$ (a constant). Prove that $E[Y] = a$.

4.4 Consider a discrete random variable X whose pmf is given by

$$f_X(x) = \begin{cases} 1/3, & x = -1, 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

Compute $E[X]$.

4.5 Let X be a uniform RV, that is,

$$f_X(x) = \begin{cases} (b-a)^{-1}, & 0 < a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

Compute $E[X]$.

4.6 Let the pdf of X be $f_X(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{else.} \end{cases}$

(i) Compute $F_X(x)$;
(ii) Compute $E[X]$;
(iii) Compute σ_X^2.

4.7 Find $E[X]$ if $P_X(x) = \dfrac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$, $x = 0, 1, ..., k$ and 0, else. This PMF is called the hypergeometric distribution and m, n, k are positive integers.

4.8 In Problem 4.5, let $Y \triangleq X^2$. Compute the pdf of Y and $E[Y]$ by Equation 4.1-8. Then compute $E[Y]$ by Equation 4.1-9.

4.9 Let $Y \triangleq X^2 + 1$. Compute $E[Y]$ and σ_Y^2 if

$$f_X(x) = \begin{cases} 2x, & 0 < x < 1, \\ 0, & \text{else.} \end{cases}$$

4.10 Let X be a Poisson RV with parameter a. Compute $E[Y]$ when $Y \triangleq X^2 + b$.

4.11 Show that the mean of the Gaussian RV $X : N(\mu, \sigma^2)$ is μ. Start from the defining integral for the mean.

4.12 In your physics courses, you have studied the concept of momentum $p = mv$ in the deterministic that is, nonrandom sense. In reality, measurements of mass m and velocity v are never precise, thereby giving rise to an unavoidable uncertainty in these quantities. In this problem, we treat these quantities as RVs. So, consider an RV mass M with given pdf $f_M(m)$ and an RV velocity V with given pdf $f_V(v)$. We are also given the averages $\mu_M = E[M]$ and $\mu_V = E[V]$ (that would presumably correspond to our measurements in the physics course). Assume that M and V are independent and nonnegative RVs.

(a) Express the pdf of the momentum $P = MV$ in terms of the known pdf's $f_M(m)$ and $f_V(v)$.
(b) Determine the expected value of the momentum $\mu_P = E[P]$ as a function of μ_M and μ_V.

4.13 Prove that if $E[X]$ exists and X is a continuous RV, then $|E[X]| \leq E[|X|]$. Repeat for X discrete.

4.14 Show that if $E[g_i(X)]$ exists for $i = 1, \ldots, N$, then

$$E\left[\sum_{i=1}^{N} g_i(X)\right] = \sum_{i=1}^{N} E[g_i(X)].$$

4.15 A random sample of 20 households shows the following numbers of children per household: 3, 2, 0, 1, 0, 0, 3, 2, 5, 0, 1, 1, 2, 0, 1, 0, 0, 0, 6, 3. (a) For this set what is the average number of children per household? (b) What is the average number of children in households given that there is at least one child?

4.16 Let $B \triangleq \{a < X \leq b\}$. Derive a general expression for $E[X|B]$ if X is a continuous RV. Let $X : N(0, 1)$ with $B = \{-1 < X \leq 2\}$. Compute $E[X|B]$.

4.17 (Papoulis [4-3]). Let $Y = h(X)$. We wish to compute approximation to $E[h(X)]$ and $E[h^2(X)]$. Assume that $h(x)$ admits to a power series expansions, that is, all derivatives exist. Assume further that all derivatives above the second are small enough to be omitted. Given that $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$, show that

(a) $E[h(X)] \simeq h(\mu) + h''(\mu)\sigma^2/2$;
(b) $E[h^2(X)] \simeq h^2(\mu) + ([h'(\mu)]^2 + h(\mu)h''(\mu))\sigma^2$.

4.18 The joint pdf of a bivariate random variable (X, Y) is given by

$$f_{XY}(x, y) = \begin{cases} 2, & 0 < y \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

(a) Find the conditional pdf of Y given $X = x$ denoted by $f_{Y/X}(y/x)$
(b) Find the conditional mean of Y given $X = x$, i.e. $E[Y/x]$
(c) Compute the mean $E[Y]$.

4.19 A particular model of an HDTV is manufactured in three different plants, say, A, B, and C, of the same company. Because the workers at A, B, and C are not equally experienced, the quality of the units differs from plant to plant. The pdf's of the time-to-failure X, in years, are

$$f_X(x) = \frac{1}{5}\exp(-x/5)u(x) \text{ for } A$$

$$f_X(x) = \frac{1}{6.5}\exp(-x/6.5)u(x) \text{ for } B$$

$$f_X(x) = \frac{1}{10}\exp(-x/10)u(x) \text{ for } C,$$

where $u(x)$ is the unit step. Plant A produces three times as many units as B, which produces twice as many as C. The TVs are all sent to a central warehouse, intermingled, and shipped to retail stores all around the country. What is the expected lifetime of a unit purchased at random?

4.20 A source transmits a signal Θ with pdf

$$f_\Theta(\theta) = \begin{cases} (2\pi)^{-1}, & 0 < \theta \leq 2\pi, \\ 0, & \text{otherwise.} \end{cases}$$

Because of additive Gaussian noise, the pdf of the received signal Y when $\Theta = \theta$ is

$$f_{Y|\Theta}(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right].$$

Compute $E[Y]$.

4.21 Compute the variance of X if X is (a) Bernoulli; (b) binomial; (c) Poisson; (d) Gaussian; (e) Rayleigh.

4.22 An Internet Service Provider (ISP) has two types of servers that route incoming packets for its customers. The servers fail randomly and have been found to have time-to-failure distributions that are exponential with parameters μ_1 and μ_2, respectively. Call these two RV failure times T_1 and T_2, and assume they are independent. Thirty percent of the servers are type 1 and 70 percent are type 2. If a server is picked at random, denote its time-to-failure by the RV T.

(a) What is $E[T]$?
(b) What is $E[T^2]$?
(c) What is the standard deviation σ_T?

4.23 Let X and Y be independent RVs, each $N(0,1)$. Find the mean and variance of $Z \triangleq \sqrt{X^2+Y^2}$.

4.24 Let X_1, X_2, X_3 be three i.i.d. standard Normal RVs. We order them as $Y_1 < Y_2 < Y_3$.

a) Compute $f_{Y_1Y_2Y_3}(y_1, y_2, y_3)$;
b) Compute $E[Y_1]$ $i = 1, 2, 3$.

4.25 Let $f_{XY}(x,y) = 2$ for $0 < x < y < 1$ and zero else. Compute $E[Y]$ and σ_Y^2.

4.26 Let $f_{XY}(x,y)$ be given by

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(-\frac{x^2+y^2-2\rho xy}{2\sigma^2(1-\rho^2)}\right),$$

where $|\rho| < 1$. Show that $E[Y] = 0$ but $E[Y|X = x] = \rho x$. What does this result say about predicting the value of Y upon observing the value of X?

4.27 Let X and Y be two Gaussian RVs with mean 0 and variance σ^2. Let

$$Z \triangleq \frac{1}{2}(X+Y).$$

(a) If X and Y are independent, what are the mean and variance of Z ?

(b) Suppose X and Y are no longer independent. Let ρ be the correlation coefficient of X and Y. Now, what would be the mean and variance of Z ? (Your answer may be in terms of ρ)?

(c) Consider what happens when $\rho = -1$, $\rho = 0$, and $\rho = +1$. Is it always true that

4.28 Show that in the joint Gaussian pdf with $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y \triangleq \sigma$, the joint pdf asymptotically as $\rho \to 1$, becomes

$$f_{XY}(x,y) \to \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right] \delta(y-x).$$

4.29 Consider a probability space $\mathscr{P} = (\Omega, \mathscr{F}, P)$. Let $\Omega = \{\zeta_1, \ldots, \zeta_5\} = \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$ with $P[\{\zeta_i\}] = \frac{1}{5}$, $i = 1, \ldots, 5$. Define two RVs on $\mathscr{P}$ as follows:

$$X(\zeta) \triangleq \zeta \qquad \text{and} \qquad Y(\zeta) \triangleq \zeta^2.$$

(a) Show that X and Y are dependent RVs.

(b) Show that X and Y are uncorrelated.

4.30 Given the conditional Gaussian density

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\alpha x)^2}{2\sigma^2}\right),$$

for two RVs X and Y, what is the conditional mean $E[Y|X]$? Here α is a known constant.

4.31 We wish to estimate the pdf of X with a function $p(x)$ that maximizes the entropy

$$H[X] \triangleq -\int_{-\infty}^{\infty} p(x) \ln p(x)\, dx.$$

It is known from measurements that $E[X] = \mu$ and $\text{Var}[X] = \sigma^2$. Find the maximum entropy estimate of the pdf of X.

4.32 Let $X : N(0, \sigma^2)$. Show that

$$m_n \triangleq E[X^n] = 1 \cdot 3 \dots (n-1)\sigma^n \qquad n \text{ even} \qquad (4.8\text{-}1)$$

$$m_n = 0 \qquad n \text{ odd}. \qquad (4.8\text{-}2)$$

4.33 With $\mu_X \triangleq E[X]$ and $\mu_Y \triangleq E[Y]$, show that if $c_{11} = \sqrt{c_{20}c_{02}}$, then

$$E\left[\left(\frac{c_{11}}{c_{20}}(X - \mu_X) - (Y - \mu_Y)\right)^2\right] = 0.$$

Use this result to show that when $|\rho| = 1$, Y is a linear function of X, that is, $Y = \alpha X + \beta$. Relate α, β to the moments of X and Y.

4.34 Show that in the optimum linear predictor in Example 4.3-4 the smallest mean-square error is

$$\varepsilon^2_{\min} = \sigma^2_Y(1 - \rho^2).$$

Explain why $\varepsilon^2_{\min} = 0$ when $|\rho| = 1$.

4.35 We are given an RV X with pdf $f_X(x) = 1 - (1/2)x$, for $0 < x < 2$ and zero else. Compute m_r, the rth moment of X for r a positive integer.

4.36 Let $E[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2$. We wish to estimate μ with the *sample mean*

$$\widehat{\mu}_N \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Compute the mean and variance of $\widehat{\mu}_N$ assuming the X_i for $i = 1, \dots, N$ are independent.

4.37 In the previous problem, how large should N be so that

$$P[|\widehat{\mu}_N - \mu| > 0.1\sigma] \leq 0.01.$$

4.38 Let X be a uniform RV in $(-\frac{1}{2}, \frac{1}{2})$. Compute (a) its moment-generating function; and (b) its mean by Equation 4.5-5. [*Hint*: sinh $z \triangleq (e^z - e^{-z})/2$. Use limits when computing the mean.]

4.39 Let X be a Poisson RV. Compute its (a) MGF; and (b) its mean by Equation 4.5-5.

4.40 The *negative binomial distribution* with parameters N, Q, P, where $Q - P = 1$, $P > 0$, and $N \geq 1$, is defined by PMF

$$P_X(k) \triangleq \binom{N+k-1}{N-1}\left(\frac{P}{Q}\right)^k\left(1 - \frac{P}{Q}\right)^N \qquad (k = 0, 1, 2, \dots).$$

It is sometimes used as an alternative to the Poisson distribution when one cannot guarantee that individual events occur independently (the "strict" randomness requirement for the Poisson distribution). Show that the moment-generating function is

$$M_X(t) = (Q - Pe^t)^{-N}.$$

[*Hint*: Either compute or look up the expansion formula for $(Q-Pe^t)^{-N}$, for example, see *Discrete Distributions* by N. L. Johnson and S. Kotz, John Wiley and Sons, 1969.]

4.41 Let X have pdf $f_X(x;\alpha,\beta) = \begin{cases} \left(\alpha!\beta^{\alpha+1}\right)^{-1} x^\alpha \exp(-x/\beta), & 0 < x < \infty, \beta > 0,\ \alpha \geq 0, \\ 0, & \text{else.} \end{cases}$
Find the moment-generating function of X. This is the *gamma distribution*.

4.42 Find the mean and variance of X if X has a gamma distribution.

4.43 Compute the Chernoff bound on $P[X \geq a]$, where X is an RV that satisfies the exponential law $f_X(x) = \lambda e^{-\lambda x}u(x)$.

4.44 Let $N = 1$ in Problem 4.40. (a) Compute the Chernoff bound on $P[X \geq k]$; (b) generalize the result for arbitrary N.

4.45 Let X have a Cauchy pdf

$$f_X(x) = \frac{\alpha}{\pi(\alpha^2 + x^2)}.$$

Compute the CF $\Phi_X(\omega)$ of X.

4.46 Let X have the Cauchy density: $f_X(x) = \left(\pi(1 + (x-a)^2)\right)^{-1}, -\infty < x < \infty$. Find $E[X]$. What problem do you run into when trying to compute σ_X^2?

4.47 Find the CF of the exponential RV X with mean $\mu > 0$, that is,

$$f_X(x) = \frac{1}{\mu}e^{-x/\mu}u(x),$$

where $u(x)$ denotes the unit-step function.

4.48 Find the characteristic function of a Cauchy random variable with pdf

$$f_X(x) = \frac{a}{\pi(x^2 + a^2)}, -\infty < x < \infty$$

If $X_1, X_2, \ldots, X_n$ are n independent Cauchy random variables with the above pdf and $Y_n = 1/n \sum_{i=1}^{n} Xi$

(a) Find the pdf of Y_n

(b) Does the Central Limit Theorem hold for Y_n?

4.49 Let X be uniform over $(-a, a)$. Let Y be independent of X and uniform over $([n-2]a, na)$, $n = 1, 2, \ldots$. Compute the expected value of $Z = X + Y$ for each n. From this result sketch the pdf of Z. What is the only effect of n?

4.50 Consider the recursion known as a *first-order moving average* given by

$$X_n = Z_n - aZ_{n-1} \qquad |a| < 1,$$

where X_n, Z_n, Z_{n-1} are all RVs for $n = \ldots, -1, 0, 1, \ldots$. Assume $E[Z_n] = 0$ all n; $E[Z_n Z_j] = 0$ all $n \neq j$; and $E[Z_n^2] = \sigma^2$ all n. Compute $R_n(k) \triangleq E[X_n X_{n-k}]$ for $k = 0, \pm 1, \pm 2, \ldots$.

4.51 Consider the recursion known as a *first-order autoregression*

$$X_n = bX_{n-1} + Z_n \qquad |b| < 1.$$

The following is assumed true: $E[Z_n] = 0$, $E[Z_n^2] = \sigma^2$ all n; $E[Z_n Z_j] = 0$ all $n \neq j$. Also $E[Z_n X_{n-j}] = 0$ for $j = 1, 2, \ldots$. Compute $R_n(k) = E[X_n X_{n-k}]$ for $k = \pm 1, \pm 2, \ldots$. Assume $E[X_n^2] \triangleq K$ independent of n.

4.52 Give an example of two random variables which are uncorrelated but not independent.

4.53 Let $f_{XY}(x, y) = 4\exp(-4[x+y])$, $x > 0, y > 0$. Find the joint MGF and CF function of (X, Y).

4.54 Let X and Y be two independent Poisson RVs with

$$P_X(k) = \frac{1}{k!}e^{-2}2^k$$

$$P_Y(k) = \frac{1}{k!}e^{-3}3^k.$$

Compute the PMF of $Z = X + Y$ using MGFs or CFs.

4.55 Your company manufactures toaster ovens. Let the probability that a toaster oven has a dent or scratch be $p = 0.05$. Assume different ovens get dented or scratched independently. In one week the company makes 2000 of these ovens. What is the approximate probability that in this week more than 110 ovens are dented or scratched?

4.56 Message length L (in bytes) on a network can be modeled as an i.i.d. exponential RV with CDF

$$P[L \leq l] \triangleq F_L(l) = \begin{cases} 1 - e^{-0.002l}, & l \geq 0, \\ 0, & l < 0. \end{cases}$$

(a) What is the expected length (in bytes) of the file necessary to store 400 messages?

(b) What is the probability that the average length of 400 randomly-chosen messages exceeds 520 bytes?

4.57 Use Chebyshev's inequality to find how many times a fair coin must be tossed in order that the probability that the ratio of the number of heads to the number of tosses will lie between 0.45 and 0.55, will be at least 0.95.

4.58 A distribution with unknown mean μ has a variance equal to 1.5. Use Central Limit Theorem to find how large a sample should be taken from the distribution in order that the probability be at least 0.95 that the sample mean will be within 0.5 of the population mean.

4.59 Let X_i for $i = 1, \ldots, n$ be a sequence of i.i.d. Bernoulli RVs with $P_X(1) = p$ and $P_X(0) = q = 1 - p$. Let the event of a $\{1\}$ be a success and the event of a $\{0\}$ be a failure.

(a) Show that

$$Z_n \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i,$$

where $W_i \triangleq (X_i - p)/\sqrt{pq}$, is a zero-mean, unity variance RV with a Normal CDF when $n >> 1$.

(b) For $n = 2000$ and $k = 110, 130, 150$ compute $P[k$ successes in n tries$]$ using (i) the exact binomial expression; (ii) the Poisson approximation to the binomial; and (iii) the CLT approximations. Do this by writing three MATLAB miniprograms. Verify that as the correct probabilities decrease, the error in the CLT approximation increases.

4.60 In Chapter 1, the following problem was solved using an approximation to the binomial probability law.

Assume that code errors in a computer program occur as follows: A line of code contains errors with probability $p = 0.001$ and is error free with probability $q = 0.999$. Also errors in different lines occur independently. In a 1000-line program, what is the approximate probability of finding 2 or more erroneous lines?

Can the Central Limit Theorem be used here to give an approximate answer? Why or why not? Explain your answer.

4.61 Assume that we have uniform random number generator (URNG) that is well modeled by a sequence of i.i.d. uniform RVs X_i, $i = 1, \ldots, n$, where X_i is the ith output of the URNG. Assume that

$$f_{X_i}(X_i) = \frac{1}{a} \text{rect}\left(\frac{x_i - a/2}{a}\right).$$

(a) Show that with $Z_n = X_1 + \ldots + X_n$, $E[Z_n] = na/2$. (b) Show that $\text{Var}(Z_n) = na^2/12$. (c) Write a MATLAB program that computes the plots $f_{Z_n}(z)$ for $n = 2, 3, 10, 20$. (d) Write a MATLAB program that plots Gaussian pdf's $N\left(\frac{na}{2}, \frac{na^2}{12}\right)$ for $n = 2, 3, 10, 20$ and compare $f_{Z_n}(z)$ with $N\left(\frac{na}{2}, \frac{na^2}{12}\right)$ for each n. (e) For each n compute $P[\mu_n - k\sigma_n \leq Z_n \leq \mu_n + k\sigma_n]$, where $\mu_n = na/2$, $\sigma_n^2 = na^2/12$ for a few values of k, for example, $k = 0.1, 0.5, 1, 2, 3$. Do this using both $f_{Z_n}(z)$ and $N\left(\frac{na}{2}, \frac{na^2}{12}\right)$. Choose any reasonable value of a, for example, $a = 1$.

4.62 Let $f_X(x)$ be the pdf of a real, continuous RV X. Show that if $f_X(x) = f_X(-x)$, then $E[X] = 0$.

4.63 Let random variables X and Y be defined by $X = \cos\Theta$ and $Y = \sin\Theta$, where Θ is a random variable uniformly distributed over $(0, 2\pi)$.

Compute $E(X)$, $E(Y)$, $E(XY)$, $E(X^2)$, $E(Y^2)$, $E(X^2Y^2)$.

4.64 Let X be a Normal RV with $X : N(\mu, \sigma^2)$. Show that $E\{(X - \mu)^{2k+1}\} = 0$, while $E[(X - \mu)^{2k}] = [(2k)!/2^k k!]\sigma^{2k}$.

4.65 (a) Write a MATLAB program (`.m` file) that will compute the pdf for a Chi-square RV Z_n and display it as a graph for $n = 30, 40, 50$. (b) Add to your program the

capability to compute $P[\mu - \sigma \leq Z_n \leq \mu + \sigma]$. Compare your result with a Gaussian approximation $P[\mu - \sigma \leq X \leq \mu + \sigma]$, where $X{:}N(n, 2n)$.

4.66 Let $X_i, i = 1, \ldots, 4$, be four zero-mean Gaussian RVs. Use the joint CF to show that

$$E[X_1X_2X_3X_4] = E[X_1X_2]E[X_3X_4] + E[X_1X_3]E[X_2X_4] + E[X_2X_3]E[X_1X_4].$$

4.67 Compute the MGF and CF for the Chi-square RV with n degrees of freedom.

4.68 Let $E[X_i] = \mu$, $\text{Var}[X_i] = \sigma^2$. We wish to estimate μ with the *sample mean*

$$\widehat{\mu} \triangleq \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Compute the mean and second moment of $\widehat{\mu}$ assuming the X_i for $i = 1, \ldots, N$ are independent.

4.69 Is the converse statement of Problem 4.62 true? That is, if $E[X] = 0$, does that imply that $f_X(x) = f_X(-x)$?

4.70 Derive the moment generating function of a random variable X with pdf $f_X(x) = \lambda e^{-\lambda x}, x > 0, \lambda > 0$ and zero otherwise. Hence obtain the mean and variance of X.

4.71 Assuming that the X_i are i.i.d. and Normal, show that $W_n \triangleq \sum_{i=1}^{n}[(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j)/\sigma]^2$ is Chi-square with $n - 1$ degrees of freedom.

4.72 (*conditional expectation*) Let $Y = X + N$, where the RVs X and N are independent Poisson RVs with means 20 and 5, respectively.

(a) Find the conditional PMF of Y given X.
(b) Find the conditional mean $E[Y|X = x]$.

4.73 Derive the inequality $\sigma_X P[|X| \geq \sigma_X] \leq E[|X|] \leq \sigma_X$ that holds true if $f_X(x) = f_X(-x)$.

4.74 Consider two RVs X and Y together with given values for μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ. We make a linear estimate of Y based on X, that is,

$$\widehat{Y} = \alpha X + \beta.$$

Define the estimate error as

$$\varepsilon \triangleq \widehat{Y} - Y.$$

(a) Then find the covariance of the estimate error and the data X, that is, find

$$\text{Cov}[\varepsilon, X] = E[\varepsilon X] - E[\varepsilon]E[X].$$

Express your answer in terms of α and β and the above given parameter values.

(b) Set α and β to their optimal values. Then evaluate $\text{Cov}[\varepsilon, X]$ again.

4.75 In Problem 4.74 we looked at estimating the RV Y from the RV X with the linear estimate

$$\widehat{Y} = \alpha X + \beta.$$

It turned out that the optimal values α and β found in class resulted in $\text{Cov}[\varepsilon, X] = 0$. Now it is relatively easy to show that this condition, that is,

$$\text{Cov}[\varepsilon, X] = 0,$$

known as *the orthogonality condition*, holds for general linear estimation problems where, as above, we want to find the best linear estimate in the sense of minimizing the mean-square error. In words we say that the estimate error ε is orthogonal to the data used in the estimate, in this case X.
Here we consider a slight generalization of this problem. We now form a linear estimate of Y based on two RVs X_1 and X_2, that is,

$$\widehat{Y} = \alpha_1 X_1 + \alpha_2 X_2 + \beta.$$

We will determine the values of α_1 and α_2 from the two orthogonality conditions

$$\text{Cov}[\varepsilon, X_1] = 0 \quad \text{and} \quad \text{Cov}[\varepsilon, X_2] = 0.$$

To make matters simpler, we assume that all three mean values are zero which implies $\beta = 0$, so that the linear estimate simplified to

$$\widehat{Y} = \alpha_1 X_1 + \alpha_2 X_2.$$

As before, the error is written as $\varepsilon = Y - \widehat{Y}$. Note that due to the means being zero, $\text{Cov}[\varepsilon, X_1] = E[\varepsilon X_1]$ and $\text{Cov}[\varepsilon, X_2] = E[\varepsilon X_2]$. Please use the following values:

$$\sigma_1^2 = 1, \sigma_2^2 = 4, \sigma_Y^2 = 4,$$
$$\rho_1 = 0.5, \rho_2 = 0.7, \rho_{12} = 0.5,$$

where $\rho_1 = E[X_1 Y]/\sigma_1\sigma_Y, \rho_2 = E[X_2 Y]/\sigma_2\sigma_Y$, and $\rho_{12} = E[X_1 X_2]/\sigma_1\sigma_2$ here since the mean values are all zero.

(a) Using these given values, write two linear equations that can be solved for α_1 and α_2 using the orthogonality conditions in the form

$$E[\varepsilon X_1] = 0 \quad \text{and} \quad E[\varepsilon X_2] = 0.$$

(b) Solve these two linear equations for α_1 and α_2.

REFERENCES

4-1. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edition. New York: John Wiley, 1971.

4-2. W. B. Davenport, Jr., *Probability and Random Processes: An Introduction for Applied Scientists and Engineers*, New York: McGraw-Hill, p. 99.
4-3. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd edition, New York: McGraw-Hill, 1991.
4-4. B. Saleh, *Photoelectron Statistics*. New York: Springer-Verlag, 1978, Chapter 5.
4-5. R. G. Gallagher, *Information Theory and Reliable Communications*. New York: John Wiley, 1968.
4-6. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Part 1. New York: McGraw-Hill, 1953, p. 279.
4-7. G. A. Korn and T. S. Korn, *Mathematical Handbook for Scientists and Engineers*. New York: McGraw-Hill, 1961.
4-8. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd edition. New York: John Wiley, 1957.

ADDITIONAL READING

Cooper, G. R. and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, 3rd edition. New York: Holt, Rinehart and Winston, 1999.

Peebles, P. Z. Jr., *Probability, Random Variables, and Random Signal Principles*, 4th edition. New York: McGraw-Hill, 2001.

Garcia, L.-G., *Probability and Random Processes for Electrical Engineering*, 2nd edition. Reading, MA: Addison-Wesley, 1994.

Helstrom, C. W., *Probability and Stochastic Processes for Engineers*, 2nd edition. New York: Macmillan, 1991.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd edition. New York: McGraw-Hill, 1991.

Scheaffer, R. L., *Introduction to Probability and Its Applications*. Belmont, CA: Duxbury, 1990.

Viniotis, Y., *Probability and Random Processes for Electrical Engineers*. New York: McGraw-Hill, 1998.

Yates, R. D. and D. J. Goodman, *Probability and Stochastic Processes*. New York: Wiley, 1999.

5 Random Vectors

5.1 JOINT DISTRIBUTION AND DENSITIES

In many practical problems involving random phenomena we make observations that are essentially of a vector nature. We illustrate with three examples.

Example 5.1-1

(*seismic discrimination*) A seismic waveform $X(t)$ is received at a geophysical recording station and is sampled at the instants $t_1, t_2, \ldots, t_n$. We thus obtain a vector $\mathbf{X} = (X_1, \ldots, X_n)^T$, where $X_i \triangleq X(t_i)$ and T denotes transpose.† For political and military reasons, at one time it was important to determine whether the waveform was radiated from an earthquake or an underground explosion. Assume that an expert computer system has available a lot of stored data regarding both earthquakes and underground explosions. The vector $\mathbf{X}$ is compared to the stored data. What is the probability that $X(t)$ is correctly identified?

Example 5.1-2

(*health vector*) To evaluate the health of grade-school children, the Health Department of a certain region measures the height, weight, blood pressure, red-blood cell count, white-blood cell count, pulmonary capacity, heart rate, blood-lead level, and vision acuity of each child. The resulting vector $\mathbf{X}$ is taken as a summary of the health of each child. What is the probability that a child chosen at random is healthy?

†All vectors will be assumed to be column vectors unless otherwise stated.

Example 5.1-3

(*disease detection*) A computer system equipped with a digital scanner is designed to recognize black-lung disease from x-rays. It does this by counting the number of radio-opacities in six lung zones (that is, three in each lung) and estimating the average size of the opacities in each zone. The result is a 12-component vector $\mathbf{X}$ from which a decision is made. What is the best computer decision?

The three previous examples are illustrative of many problems encountered in engineering and science that involve a number of random variables (RVs) that are grouped for some purpose. Such groups of RVs are conveniently studied by vector methods. For this reason we treat these grouped RVs as a single object called a *random vector*. As in earlier chapters, capital letters at the lower end of the alphabet will denote RVs; bold capital letters will denote random vectors and matrices and lowercase bold letters are deterministic vectors, for example, the values that random vectors assume.

Consider a sample description space Ω with point ζ and a set of n real RVs $X_1, X_2, \cdots, X_n$ from Ω to the real line R. For each $\zeta \in \Omega$ we generate the n-component vector of numbers $\mathbf{X}(\zeta) \triangleq (X_1(\zeta), X_2(\zeta), \ldots, X_n(\zeta)) \in R^n$. Then $\mathbf{X} \triangleq (X_1, X_2, \ldots, X_n)$ is said to be an n-dimensional real random vector. The definition is readily extended to a complex random vector. Let $\mathbf{X}$ be an n-dimensional random vector defined on sample space Ω with CDF $F_{\mathbf{X}}(\mathbf{x})$. Then by definition[†]

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq P[X_1 \leq x_1, \ldots, X_n \leq x_n]. \tag{5.1-1}$$

By defining $\{\mathbf{X} \leq \mathbf{x}\} \triangleq \{X_1 \leq x_1, \ldots, X_n \leq x_n\}$, we can rewrite Equation 5.1-1 concisely as

$$F_{\mathbf{X}}(\mathbf{x}) \triangleq P[\mathbf{X} \leq \mathbf{x}]. \tag{5.1-2}$$

We associate the events $\{\mathbf{X} \leq \infty\}$ and $\{\mathbf{X} \leq -\infty\}$ with the certain event Ω and impossible event ϕ, respectively. Hence

$$F_{\mathbf{X}}(\infty) = 1 \tag{5.1-3a}$$

$$F_{\mathbf{X}}(-\infty) = 0. \tag{5.1-3b}$$

If the nth-mixed partial of $F_{\mathbf{X}}(\mathbf{x})$ exists we can define a *probability density function* (pdf) as

$$f_{\mathbf{X}}(\mathbf{x}) \triangleq \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \ldots \partial x_n}. \tag{5.1-4}$$

[†]We remind the reader that the event $\{X_1 \leq x_1, \ldots, X_n \leq x_n\}$ is the *intersection* of the n events $\{X_i \leq x_i\}$ for $i = 1, \ldots n$. If any one of these sub-events is the impossible event e.g., $\{X_i \leq -\infty\}$ then the the whole event becomes the impossible event and we would still write $F_X(-\infty) = 0$.

The reader will observe that these definitions are completely analogous to the scalar definitions given in Chapter 2. We could have defined

$$f_{\mathbf{X}}(\mathbf{x}) \triangleq \lim_{\substack{\Delta x_1 \to 0 \\ \vdots \\ \Delta x_n \to 0}} \frac{P[x_1 < X_1 \le x_1 + \Delta x_1, \ldots, x_n < X_n \le x_n + \Delta x_n]}{\Delta x_1 \ldots \Delta x_n} \tag{5.1-5}$$

and arrived at Equation 5.1-4. For example, for $n = 2$

$$\begin{aligned} &P[x_1 < X_1 \le x_1 + \Delta x_1, x_2 < X_2 \le x_2 + \Delta x_2] \\ &\quad = F_{\mathbf{X}}(x_1 + \Delta x_1, x_2 + \Delta x_2) - F_{\mathbf{X}}(x_1, x_2 + \Delta x_2) - F_{\mathbf{X}}(x_1 + \Delta x_1, x_2) + F_{\mathbf{X}}(x_1, x_2). \end{aligned}$$

Thus (still for $n = 2$)

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) = \lim_{\substack{\Delta x_1 \to 0 \\ \Delta x_2 \to 0}} \frac{1}{\Delta x_1 \Delta x_2} &[F_{\mathbf{X}}(x_1 + \Delta x_1, x_2 + \Delta x_2) - F_{\mathbf{X}}(x_1 + \Delta x_1, x_2) \\ &- F_{\mathbf{X}}(x_1, x_2 + \Delta x_2) + F_{\mathbf{X}}(x_1, x_2)] \end{aligned}$$

which is by definition the second mixed partial derivative, and thus

$$f_{\mathbf{X}}(x_1, x_2) = \frac{\partial^2 F_{\mathbf{X}}(x_1, x_2)}{\partial x_1 \partial x_2}.$$

From Equation 5.1-5 we make the useful observation that

$$f_{\mathbf{X}}(\mathbf{x}) \Delta x_1 \ldots \Delta x_n \simeq P[x_1 < X_1 \le x_1 + \Delta x_1, \ldots, x_n < X_n \le x_n + \Delta x_n] \tag{5.1-6}$$

if the increments are small. If we integrate Equation 5.1-4, we obtain the CDF as

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(\mathbf{x}')\, dx_1' \ldots dx_n',$$

which we can write in compact notation as

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_{\mathbf{X}}(\mathbf{x}') d\mathbf{x}'.$$

More generally, for any event $B \subset R^N$ (R^N being Euclidean N-space) consisting of the countable union and intersection of parallelepipeds

$$P[B] = \int_{\mathbf{x} \in B} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \tag{5.1-7}$$

(Compare with Equation 2.5-3.) The argument behind the validity of Equation 5.1-7 follows very closely the argument furnished in the one-dimensional case (Section 2.5). Davenport [5-1, p. 149] discusses the validity of Equation 5.1-7 for the case $n = 2$. For $n > 2$ one can proceed by induction.

The CDF of $\mathbf{X}$ given the event B is defined by

$$F_{\mathbf{X}|B}(\mathbf{x}|B) \triangleq P[\mathbf{X} \leq \mathbf{x}|B]$$
$$= \frac{P[\mathbf{X} \leq \mathbf{x}, B]}{P[B]} \quad (P[B] \neq 0).$$

These and subsequent results closely parallel the one-dimensional case. Consider next the n disjoint and exhaustive events $\{B_i, i = 1, \ldots, n\}$ with $P[B_i] > 0$. Then $\cup_{i=1}^n B_i = \Omega$ and $B_i B_j = \phi$ for all $i \neq j$. From the Total Probability Theorem 1.6-1, it then follows that

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{n} F_{\mathbf{X}|B_i}(\mathbf{x}|B_i)P[B_i]. \tag{5.1-8}$$

The unconditional CDF on the left is sometimes called a *mixture* distribution function. The conditional pdf of $\mathbf{X}$ given the event B is an nth mixed partial derivative of $F_{\mathbf{X}|B}(\mathbf{x}|B)$ if it exists. Thus,

$$f_{\mathbf{X}|B}(\mathbf{x}|B) \triangleq \frac{\partial^n F_{\mathbf{X}|B}(\mathbf{x}|B)}{\partial x_1 \ldots \partial x_n}. \tag{5.1-9}$$

It follows from Equations 5.1-8 and 5.1-9 that

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{n} f_{\mathbf{X}|B}(\mathbf{x}|B_i)P[B_i]. \tag{5.1-10}$$

Because $f_{\mathbf{X}}(\mathbf{x})$ is a mixture, that is, a linear combination of conditional pdf's, it is sometimes called a *mixture* pdf.†

The joint CDF of two random vectors $\mathbf{X} = (X_1, \ldots, X_n)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)^T$ is

$$F_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = P[\mathbf{X} \leq \mathbf{x}, \mathbf{Y} \leq \mathbf{y}]. \tag{5.1-11}$$

The *joint density* of $\mathbf{X}$ and $\mathbf{Y}$, if it exists, is given by

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = \frac{\partial^{(n+m)} F_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})}{\partial x_1 \ldots \partial x_n \, \partial y_1 \ldots \partial y_m}. \tag{5.1-12}$$

The *marginal density* of $\mathbf{X}$ alone, $f_{\mathbf{X}}(\mathbf{x})$, can be obtained from $f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y})$ by integration, that is,

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \, dy_1 \ldots dy_m.$$

Similarly, the marginal pdf of a reduced vector $\mathbf{X}' \triangleq (X_1, \ldots, X_{n-1})^T$ is obtained from the pdf of $\mathbf{X}$ by

$$f_{\mathbf{X}'}(\mathbf{x}') \triangleq \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) \, dx_n \quad \text{where } \mathbf{x}' \triangleq (x_1, \ldots, x_{n-1})^T. \tag{5.1-13}$$

Obviously, Equation 5.1-13 can be extended to all the other marginal pdf's as well by merely integrating over the appropriate variable.

†This usage is prevalent in statistical pattern recognition.

Example 5.1-4

(*particle at random*) Let $\mathbf{X} = (X_1, X_2, X_3)^T$ denote the position of a particle inside a sphere of radius a centered about the origin. Assume that at the instant of observation, the particle is equally likely to be anywhere in the sphere, that is,

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \dfrac{3}{4\pi a^3}, & \sqrt{x_1^2 + x_2^2 + x_3^2} < a, \\ 0, & \text{otherwise.} \end{cases}$$

Compute the probability that the particle lies within a subsphere of radius $2a/3$ contained within the larger sphere.

Solution Let E denote the event that the particle lies within the subsphere (centered at the origin for simplicity) and let

$$\mathscr{R} \triangleq \{x_1, x_2, x_3 \colon \sqrt{x_1^2 + x_2^2 + x_3^2} < 2a/3\}.$$

Then the evaluation of

$$P[E] = \iiint f_{\mathbf{x}}(x_1, x_2, x_3)\, dx_1\ dx_2\ dx_3$$

is best done using spherical coordinates, that is,

$$P[E] = \frac{3}{4\pi a^3} \int_{r=0}^{2a/3} \int_{\phi=0}^{\pi} \int_{\theta=0}^{2\pi} r^2 \sin\phi\, dr\, d\phi\, d\theta.$$

Note that in this simple case the answer can be obtained directly by noting the ratio of volumes, that is, $(2a/3)^3 \div a^3 = 8/27 \simeq 0.3$.

5.2 MULTIPLE TRANSFORMATION OF RANDOM VARIABLES

The material in this section is a direct extension of Section 3.4 in Chapter 3. Let $\mathbf{X}$ be an n-dimensional random vector defined on sample space Ω. Then consider the n real functions

$$\begin{aligned} y_1 &= g_1(x_1, x_2, \cdots, x_n) \\ y_2 &= g_2(x_1, x_2, \cdots, x_n) \\ &\vdots \\ y_n &= g_n(x_1, x_2, \cdots, x_n), \end{aligned} \tag{5.2-1}$$

where the $g_i, i = 1, \ldots, n$ are *functionally independent*, meaning that there exists no function $H(y_1, y_2, \ldots, y_n)$ that is identically zero. For example, the three linear functions

$$\begin{aligned} y_1 &= x_1 - 2x_2 + x_3 \\ y_2 &= 3x_1 + 2x_2 + 2x_3 \\ y_3 &= 5x_1 - 2x_2 + 4x_3 \end{aligned} \tag{5.2-2}$$

are not functionally independent because $H(y_1, y_2, \ldots, y_n) = 2y_1 + y_2 - y_3 = 0$ for all values of x_1, x_2, x_3. We create the vector of n RVs $\mathbf{Y} \triangleq (Y_1, Y_2, \ldots, Y_n)$ according to

$$\begin{aligned} Y_1 &= g_1(X_1, X_2, \cdots, X_n) \\ Y_2 &= g_2(X_1, X_2, \cdots, X_n) \\ &\vdots \\ Y_n &= g_n(X_1, X_2, \cdots, X_n). \end{aligned} \tag{5.2-3}$$

In this way we have generated n functions of n RVs. In order to save on notation, we let $\mathbf{x} \triangleq (x_1, x_2, \ldots, x_n)$, $\mathbf{y} \triangleq (y_1, y_2, \ldots, y_n)$ and ask: Given the joint pdf $f_{\mathbf{X}}(\mathbf{x})$, how do we compute the joint pdf of the $Y_i, i = 1, \ldots, n$, that is $f_{\mathbf{Y}}(\mathbf{y})$? Note that if we start out with fewer RVs Y_i, say $i = 1, \ldots, m$, than the number of X_i, say $i = 1, \ldots, n$ with $m < n$, we can add more Y_i by introducing auxiliary functions as we did in Example 3.5-4.

We assume that we can solve the set of Equations 5.2-1 uniquely for the $x_i, i = 1, \ldots, n$, as

$$\begin{aligned} x_1 &= \phi_1(y_1, y_2, \cdots, y_n) \\ x_2 &= \phi_2(y_1, y_2, \cdots, y_n) \\ &\vdots \\ x_n &= \phi_n(y_1, y_2, \cdots, y_n). \end{aligned} \tag{5.2-4}$$

Now consider the infinitessimal event $A \triangleq \{\zeta\colon y_i < Y_i \leq y_i + dy_i,\ i = 1, \ldots, n\}$. Here the Y_i are restricted to take on values in the infinitesimal *rectangular* parallelepiped that we denote by $\mathscr{P}_y$. Following the procedure in Equations 3.4-5 to 3.4-8, we write

$$P[A] = \int_{\mathscr{P}_y} f_{\mathbf{Y}}(\mathbf{y})\, dy = f_{\mathbf{Y}}(\mathbf{y}) \mathrm{V}_y = \int_{\mathscr{P}_x} f_{\mathbf{X}}(\mathbf{x})\, dx = f_{\mathbf{X}}(\mathbf{x}) \mathrm{V}_x, \tag{5.2-5}$$

where $\mathscr{P}_x$ is an infinitesimal parallelepiped (not necessarily rectangular), V_y is the volume of $\mathscr{P}_y$, and V_x is the volume of $\mathscr{P}_x$. From Equation 5.2-5 we obtain

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \frac{V_x}{V_y}. \tag{5.2-6}$$

The ratio of infinitesimal volumes is shown in Appendix C to be the magnitude of the determinant $\tilde{J}$, given by

$$\tilde{J} = \begin{vmatrix} \dfrac{\partial \phi_1}{\partial y_1} & \cdots & \dfrac{\partial \phi_1}{\partial y_n} \\ \vdots & & \vdots \\ \dfrac{\partial \phi_n}{\partial y_1} & \cdots & \dfrac{\partial \phi_n}{\partial y_n} \end{vmatrix} \tag{5.2-7}$$

$$= \begin{vmatrix} \dfrac{\partial g_1}{\partial x_1} & \cdots & \dfrac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial g_n}{\partial x_1} & \cdots & \dfrac{\partial g_n}{\partial x_n} \end{vmatrix}^{-1} = J^{-1}. \tag{5.2-8}$$

Hence

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})|\tilde{J}| = f_{\mathbf{X}}(\mathbf{x})/|J|. \tag{5.2-9}$$

In general, the infinitesimal rectangular parallelepiped in the $\mathbf{y}$ system maps into r disjoint, infinitesimal parallelepipeds in the $\mathbf{x}$ system. Then the event A, as defined above, is the union of the events $E_i, i = l, \ldots, r$, where $E_i = \{\mathbf{X} \in \mathscr{P}_x^{(i)}\}$ and $\mathscr{P}_x^{(i)}$ is one of the r parallelepipeds in the $\mathbf{x}$ system with volume $V_x^{(i)}$. Since the regions and, therefore, the events are disjoint, the elementary probabilities $P[E_i]$ add, and we obtain the main result of this section, that is,

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^{r} f_{\mathbf{X}}(\mathbf{x}^{(i)})|\tilde{J}_i| \tag{5.2-10}$$

$$= \sum_{i=1}^{r} f_{\mathbf{X}}(\mathbf{x}^{(i)})/|J_i|. \tag{5.2-11}$$

In Equations 5.2-10 and 5.2-11 $|\tilde{J}_i| \triangleq V_x^{(i)}/V_y$ and $|\tilde{J}_i| = |J_i|^{-1}$.

Example 5.2-1

(*vector transformation*) We are given three scalar transformations of vector $\mathbf{x}$

$$g_1(\mathbf{x}) = x_1^2 - x_2^2$$
$$g_2(\mathbf{x}) = x_1^2 + x_2^2$$
$$g_3(\mathbf{x}) = x_3.$$

There are four solutions (roots) to the system

$$y_1 = x_1^2 - x_2^2$$
$$y_2 = x_1^2 + x_2^2$$
$$y_3 = x_3.$$

They are

$$\begin{aligned}
x_1^{(1)} &= ((y_1 + y_2)/2)^{1/2} & x_1^{(2)} &= ((y_1 + y_2)/2)^{1/2} \\
x_2^{(1)} &= ((y_2 - y_1)/2)^{1/2} & x_2^{(2)} &= -((y_2 - y_1)/2)^{1/2} \\
x_3^{(1)} &= y_3 & x_3^{(2)} &= y_3 \\
x_1^{(3)} &= -((y_1 + y_2)/2)^{1/2} & x_1^{(4)} &= -((y_1 + y_2)/2)^{1/2} \\
x_2^{(3)} &= ((y_2 - y_1)/2)^{1/2} & x_2^{(4)} &= -((y_2 - y_1)/2)^{1/2} \\
x_3^{(3)} &= y_3 & x_3^{(4)} &= y_3.
\end{aligned} \tag{5.2-12}$$

For the roots to be real, $y_2 \geq 0$, $y_1 + y_2 \geq 0$, and $y_2 - y_1 \geq 0$. Hence $y_2 \geq |y_1|$. In this case the single rectangular parallelepiped in the three-dimensional $\mathbf{y}$ space maps into four disjoint, infinitesimal parallelepipeds in three-dimensional $\mathbf{x}$ space.

Example 5.2-2

(*more vector transformation*) For the transformation considered in Example 5.2-1, compute $f_{\mathbf{Y}}(\mathbf{y})$ if

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-3/2}\exp\left[-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right],$$

i.e. $\mathbf{X}$ is a three-dimensional standard Gaussian RV.

Solution We must compute the Jacobian $|J|$ at each of the four roots. The Jacobian is computed as

$$J = \begin{vmatrix} 2x_1 & -2x_2 & 0 \\ 2x_1 & +2x_2 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 8x_1x_2.$$

For example at the first root we compute

$$J_1 = 4(y_2^2 - y_1^2)^{1/2}.$$

A direct calculation shows that $|J_1| = |J_2| = |J_3| = |J_4|$. Finally labeling the four solutions in Equation 5.2-12 as $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, we obtain

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{4(y_2^2 - y_1^2)^{1/2}} \sum_{i=1}^{4} f_{\mathbf{X}}(\mathbf{x}_i) \\ &= \frac{(2\pi)^{-3/2}}{(y_2^2 - y_1^2)^{1/2}} \exp\left[-\frac{1}{2}(y_2 + y_3^2)\right] \times u(y_2)u(y_2 - |y_1|)^{\dagger}. \end{aligned}$$

Although a random vector is completely characterized by its distribution or density function, the latter is often hard to come by except for some notable exceptions. By far the two most important exceptions are (1) when $F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1)\ldots F_{X_n}(x_n)$, that is, the n components of $\mathbf{X}$ are independent, and (2) when $\mathbf{X}$ obeys the multidimensional Gaussian law. Case (1) is easily handled, since it is a direct extension of the scalar case. Case (2) will be discussed in Section 5.6. But what to do when neither case (1) nor (2) applies? Estimating multidimensional distributions involving dependent variables is often not practical and even if available might be too complex to be of any real use. Therefore, when we deal with vector RVs, we often settle for a less complete but more computable characterization based on moments. For most engineering applications, the most important moments are the expectation vector (the first moment) and the covariance matrix (a second moment). These quantities and their use are discussed later on in the chapter. Next, we consider random vectors with ordered components.

5.3 ORDERED RANDOM VARIABLES

In Section 3.4 (Examples 3.4-3 and 3.4-5) we introduced the notion of two ordered RVs. Here we generalize to n RVs and obtain some important results regarding these. Ordered RVs are quite important because in the absence of any information about the distribution of the RVs, the statistics of the ordering transformation can give us significant information about such parameters as the median, range, and others that are closely related to the

† It would be challenging to show that this pdf integrates to unity.

parameters of distributions. Consider n i.i.d. *continuous* RVs each with pdf $f_X(x)$, where $-\infty < x < \infty$. The joint pdf of all n RVs is $f_{X_1 \cdots X_n}(x_1, \cdots, x_n) \doteq f_X(x_1) \cdots f_X(x_n)$ and the joint marginal density of, say, X_1 and X_n is obtained by integrating out with respect to $x_2, ..., x_{n-1}$. Now arrange the n RVs in order of increasing size; that is, if $X_k = \min(X_1, \cdots, X_n)$ then $Y_1 = X_k$, and Y_2 is the next smallest of the $\{X_i, i = 1, ..., n, i \neq k\}$, and Y_3 is the next smallest after that until, finally, $Y_n = \max(X_1, \cdots, X_n)$. We thus have performed an ordering transformation, and we can write that the strict inequalities $Y_1 < Y_2 < \cdots < Y_{n-1} < Y_n$ occur with probability 1, since the X_i are assumed continuous RVs. We wish to find the joint pdf of the $\{Y_i, i = 1, ..., n\}$. At first glance we might argue, incorrectly, that since the set $S_1 = \{X_i, i = 1, .., n\}$ contains the same elements as the set $S_2 = \{Y_i, i = 1, .., n\}$,

$$\begin{aligned} f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) &= f_{X_1 \cdots X_n}(y_1, \cdots, y_n) \\ &= f_{X_1}(y_1) \cdots f_{X_n}(y_n) \end{aligned}$$

for $\{y_i : -\infty < y_i < \infty, i = 1, ..., n\}$. However, this result ignores the fact that the $\{Y_i, i = 1, .., n\}$ *are not independent random variables.* For example if you have observed X_1, what have you learned about X_2 from observing X_1? Nothing it turns out but if you are given Y_1, you know right away that $Y_2 > Y_1$ and you also know that the probability that $Y_1 > Y_2$ is zero. Hence there is no probability mass in the region $y_1 > y_2$. With this in mind we might want to modify the joint pdf's of the $\{Y_i\}$ to

$$\begin{aligned} f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) &= f_{X_1}(y_1) \cdots f_{X_n}(y_n) \quad \text{for } y_1 < y_2 < \cdots y_n \\ &= 0, \text{ else.} \end{aligned}$$

However, now we have another problem: The volume enclosed by the modified joint pdf is not unity. Indeed for n large it could be substantially smaller than unity. To get the correct joint pdf for the $\{Y_i\}$, we shall use the results of Section 5.2, which allow us to compute the pdf of one set of RVs that are functionally related to another set whose pdf we already know.

We begin by partitioning the n-dimensional space $(-\infty < x_1, x_2, \cdots, x_n < \infty)$ into $n!$ nonoverlapping, distinct regions described by $\mathscr{R}_i = \{x_{i(1)} < x_{i(2)} < \cdots x_{i(j)} < \cdots < x_{i(n)}\}$ for $1 \leq i \leq n!$, $1 \leq j \leq n$, and $x_{i(j)} \in \{x_1, x_2, \cdots, x_n\}$. Note that $x_{i(j)} < x_{i(k)}$ for $j < k$. Each region will have a different size-ordering of its elements. For example, consider 3-space (x_1, x_2, x_3). Then a distinct, nonoverlapping partition is

$$\begin{aligned} \mathscr{R}_1 &= (x_1 < x_2 < x_3) \\ \mathscr{R}_2 &= (x_1 < x_3 < x_2) \\ \mathscr{R}_3 &= (x_2 < x_1 < x_3) \\ \mathscr{R}_4 &= (x_2 < x_3 < x_1) \\ \mathscr{R}_5 &= (x_3 < x_1 < x_2) \\ \mathscr{R}_6 &= (x_3 < x_2 < x_1). \end{aligned}$$

For each of the $n!$ regions we define $y_1 \triangleq x_{j(1)} < y_2 \triangleq x_{j(2)} < \cdots y_n \triangleq x_{j(n)}; j = 1, ..., n!$. For example in 3-space (x_1, x_2, x_3) we have

$$\begin{aligned} &\text{for } \mathscr{R}_1 : y_1 = x_1; y_2 = x_2; y_3 = x_3 \\ &\text{for } \mathscr{R}_2 : y_1 = x_1; y_2 = x_3; y_3 = x_3 \\ &\text{for } \mathscr{R}_3 : y_1 = x_2; y_2 = x_1; y_3 = x_3 \\ &\text{for } \mathscr{R}_4 : y_1 = x_2; y_2 = x_3; y_3 = x_1 \\ &\text{for } \mathscr{R}_5 : y_1 = x_3; y_2 = x_1; y_3 = x_2 \\ &\text{for } \mathscr{R}_6 : y_1 = x_3; y_2 = x_2; y_3 = x_1. \end{aligned}$$

Thus in 3-space (x_1, x_2, x_3), there are six sets of transformation equations and $6 = 3!$ distinct solutions for $y_1 < y_2 < y_3$; there are no solutions in y-space otherwise:

$$
\begin{aligned}
\text{in } \mathscr{R}_1 : y_1 &= g_1(x_1,x_2,x_3) = x_1; x_1^{(1)} = \phi_1(y_1,y_2,y_3) = y_1\\
y_2 &= h_1(x_1,x_2,x_3) = x_2; x_2^{(1)} = \varphi_1(y_1,y_2,y_3) = y_2\\
y_3 &= q_1(x_1,x_2,x_3) = x_3; x_3^{(1)} = \theta_1(y_1,y_2,y_3) = y_3\\
\text{in } \mathscr{R}_4 : y_1 &= g_4(x_1,x_2,x_3) = x_2; x_2^{(4)} = \phi_4(y_1,y_2,y_3 = y_1\\
y_2 &= h_4(x_1,x_2,x_3) = x_3; x_3^{(4)} = \varphi_4(y_1,y_2,y_3) = y_2\\
y_3 &= q_4(x_1,x_2,x_3) = x_1; x_1^{(4)} = \theta_4(y_1,y_2,y_3) = y_3\\
\text{in } \mathscr{R}_2 : y_1 &= g_2(x_1,x_2,x_3) = x_1; x_1^{(2)} = \phi_2(y_1,y_2,y_3 = y_1\\
y_2 &= h_2(x_1,x_2,x_3) = x_3; x_3^{(2)} = \varphi_2(y_1,y_2,y_3) = y_2\\
y_3 &= q_2(x_1,x_2,x_3) = x_2; x_2^{(2)} = \theta_2(y_1,y_2,y_3) = y_3\\
\text{in } \mathscr{R}_5 : y_1 &= g_5(x_1,x_2,x_3) = x_3; x_3^{(5)} = \phi_5(y_1,y_2,y_3 = y_1\\
y_2 &= h_5(x_1,x_2,x_3) = x_1; x_1^{(5)} = \varphi_5(y_1,y_2,y_3) = y_2\\
y_3 &= q_5(x_1,x_2,x_3) = x_2; x_2^{(5)} = \theta_5(y_1,y_2,y_3) = y_3\\
\text{in } \mathscr{R}_3 : y_1 &= g_3(x_1,x_2,x_3) = x_2; x_2^{(3)} = \phi_3(y_1,y_2,y_3 = y_1\\
y_2 &= h_3(x_1,x_2,x_3) = x_1; x_1^{(3)} = \varphi_3(y_1,y_2,y_3) = y_2\\
y_3 &= q_3(x_1,x_2,x_3) = x_3; x_3^{(3)} = \theta_3(y_1,y_2,y_3) = y_3\\
\text{in } \mathscr{R}_6 : y_1 &= g_6(x_1,x_2,x_3) = x_3; x_3^{(6)} = \phi_6(y_1,y_2,y_3 = y_1\\
y_2 &= h_6(x_1,x_2,x_3) = x_2; x_2^{(6)} = \varphi_6(y_1,y_2,y_3) = y_2\\
y_3 &= q_6(x_1,x_2,x_3) = x_1; x_1^{(6)} = \theta_6(y_1,y_2,y_3) = y_3.
\end{aligned}
$$

The magnitude of the Jacobian of each of these transformations is unity so that Equation 5.2-11, specialized here (in slightly different notation) for three ordered i.i.d. RVs, yields

$$
\begin{aligned}
f_{Y_1Y_2Y_3}(y_1,y_2,y_3) &= \sum_{m=1}^{m=6} \frac{f_{X_1X_2X_3}(x_1^{(m)}, x_2^{(m)}, x_3^{(m)})}{|J_m|}\\
&= \sum_{m=1}^{m=6} f_X(x_1^{(m)}) f_X(x_2^{(m)}) f_X(x_3^{m)}).
\end{aligned}
$$

Finally, expanding the summation and inserting the appropriate solutions, we obtain

$$
\begin{aligned}
f_{Y_1Y_2Y_3}(y_1,y_2,y_3) &= f_X(y_1)f_X(y_2)f_X(y_3) + f_X(y_1)f_X(y_3)f_X(y_2) + f_X(y_2)f_X(y_1)f_X(y_3)\\
&\quad + f_X(y_2)f_X(y_3)f_X(y_1) + f_X(y_3)f_X(y_1)f_X(y_2) + f_X(y_3)f_X(y_2)f_X(y_1)\\
&= 3! f_X(y_1)f_X(y_2)f_X(y_3).
\end{aligned}
$$

This result applies when $y_1 < y_2 < y_3$; otherwise $f_{Y_1Y_2Y_3}(y_1, y_2, y_3) = 0$.

We now summarize the result for the general case. We are given n continuous i.i.d. RVs with pdf $f_{X_1 \cdots X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_X(x_i)$ with $-\infty < x_1, x_2, \cdots, x_n < \infty$ and consider the transformation that orders them by signed magnitude so that $Y_1 < Y_2 < \cdots < Y_n$, where for $i = 1, ..., n$, $Y_i \in \{X_1, X_2, \cdots, X_n\}$. Then

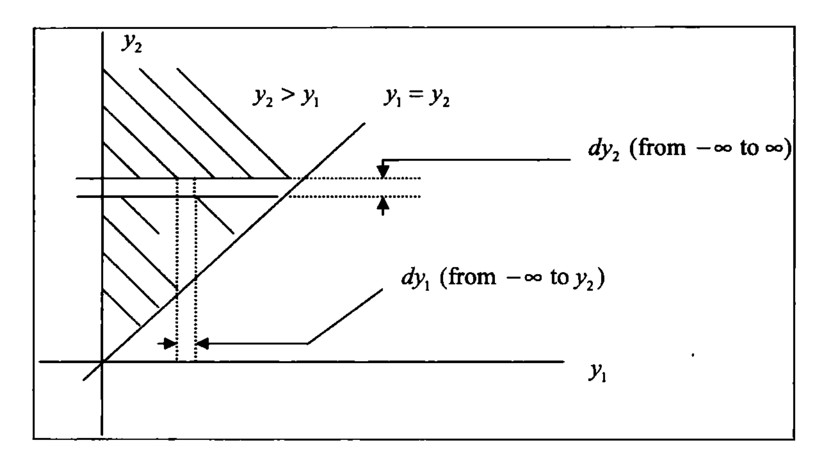


Figure 5.3-1 Showing integration regions for two ordered random variables.

$$f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^{n} f_X(y_i), & \text{for } -\infty < y_1 < y_2 < \cdots < y_n < \infty \\ 0, & \text{else.} \end{cases} \tag{5.3-1}$$

If $f_{Y_1 \cdots Y_n}(y_1, \cdots, y_n)$ is a true pdf, it must integrate out to 1. This requires an n-fold iterated integration.

To show how this integration is done we consider the $n = 2$ case. Then integrating the function $2f_{Y_1Y_2}(y_1, y_2) = 2f_{Y_1}(y_1)f_{Y_2}(y_2)$ over the region $-\infty < y_1 < y_2 < \infty$ requires integrating the integrand from $-\infty < y_1 < y_2$ followed by an integration from $-\infty < y_2 < \infty$. This is shown in Figure 5.3-1. Since $-\infty < y_1 < y_2$ we integrate the y_1 variable from $-\infty$ to y_2; then we complete the integration over the half-space by integrating the y_2 variable from $-\infty$ to ∞.

The extension to the n-dimensional case is straightforward: We integrate the y_1 variable first from $-\infty$ to y_2; next the y_2 variable from $-\infty$ to y_3, etc.; finally the y_n variable gets integrated from $-\infty$ to ∞. In this fashion we have integrated over the entire subspace $-\infty < y_1 < \cdots < y_n < \infty$. The last integration yields

$$n! \int_{-\infty}^{\infty} F_X^{n-1}(y_n) f_X(y_n) dy_n/(n-1)! = n! \int_{-\infty}^{\infty} F_X^{n-1}(y_n) dF(y_n)/(n-1)! = F_X^n(y_n)|_{-\infty}^{\infty} = 1.$$

The next development leads to the fundamental result of order statistics.

Distribution of area random variables

We begin by defining the area RVs

$$Z_i \triangleq \int_{-\infty}^{Y_i} f_X(x)dx, i = 1, ..., n, \tag{5.3-2}$$

where $f_X(x)$ is the pdf of a continuous RV X, and $X_1, \cdots, X_n$ are n i.i.d. observations on X. After ordering we obtain the $Y_i, i = 1, ..., n$, as the ordered RVs where $\min(X_1, ..., X_n) \triangleq Y_1 < Y_2 < \cdots < Y_n \triangleq \max(X_1, ..., X_n)$. We denote Z_i somewhat informally as an "area RV" because the RV Z_i is the area under $f_X(x)$ up to Y_i. Clearly, because Y_i is an RV so is Z_i. Indeed, we can think of Z_i as a CDF with a random argument, hence we may also speak of it as a random CDF. We recognize that $Z_1 < \cdots < Z_n$ because $Y_1 < \cdots < Y_n$ and Z_i is a monotonically increasing function of Y_i for every index i. We consider the transformation

$$z_i = \int_{-\infty}^{y_i} f_X(x)dx = F_X(y_i), i = 1, ..., n,$$

where $F_X(x)$ is a continuously increasing function of x, and hence has a unique inverse at every x. The roots of these equations are $y_i^{(r)} = F_X^{-1}(z_i), i = 1, ..., n,$ (see Figure 5.3-2) and the Jacobian is

$$\begin{vmatrix} \frac{\partial z_1}{\partial y_1} & \bullet\bullet\bullet & \frac{\partial z_1}{\partial y_n} \\ \bullet & \bullet & \\ \bullet & \bullet & \\ \bullet & \bullet & \\ \frac{\partial z_n}{\partial y_1} & & \frac{\partial z_n}{\partial y_n} \end{vmatrix} = \begin{vmatrix} f_X(y_1^{(r)}) & 0 & \bullet & \bullet & 0 \\ 0 & f_X(y_2^{(r)}) & 0 & & \bullet \\ \bullet & 0 & \ddots & & \bullet \\ \bullet & & & \ddots & 0 \\ 0 & \bullet & \bullet & 0 & f_X(y_n^{(r)}) \end{vmatrix} = \prod_{i=1}^{n} f_X(y_i^{(r)}). \tag{5.3-3}$$

Hence the pdf of the $Z_i, i = 1, ..., n$, is determined as

$$\begin{aligned} f_{Z_1 \cdots Z_n}(z_1, \cdots, z_n) &= n! \frac{\prod_{i=1}^{n} f_X(y_i^{(r)})}{\prod_{i=1}^{n} f_X(y_i^{(r)})} = n!, \; 0 < z_1 < z_2 < \cdots < z_n < 1 \\ &= 0, \text{ else.} \end{aligned} \tag{5.3-4}$$

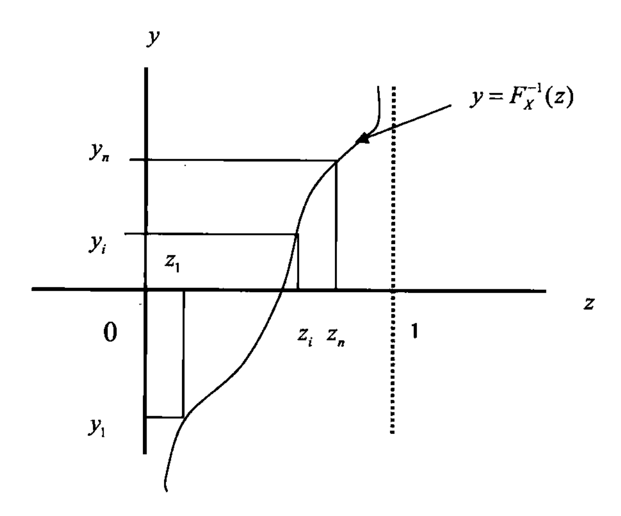


Figure 5.3-2 Finding the roots of the transformation $y = F_X^{-1}(z)$.

This non-intuitive result says that the pdf of the $Z_i, i = 1, ..., n$, *does not depend on the underlying pdf* $f_X(x)$. Equation 5.3-4 enables us to derive a number of important results useful in estimating various parameters when we don't know the underlying distributions. See for Example 5.3-5.

Example 5.3-1

(*area under* $f_X(x)$ *between the smallest and largest observations*) We wish to compute the area under $f_X(x)$ between the smallest, $Y_1 = \min(X_1, \cdots, X_n)$, and largest, $Y_n = \max(X_1, \cdots, X_n)$, of the observations in a sample of size n drawn from the pdf $f_X(x)$. We denote this area with the new random variable

$$V_{1n} \triangleq \int_{Y_1}^{Y_n} f_X(x)dx \,. \tag{5.3-5}$$

We note that

$$V_{1n} = \int_{-\infty}^{Y_n} f_X(x)dx - \int_{-\infty}^{Y_1} f_X(x)dx = Z_n - Z_1 \tag{5.3-6}$$

hence we need to compute $f_{Z_1 Z_n}(z_1, z_n)$ from $f_{Z_1 \cdots Z_n}(z_1, \cdots, z_n)$. This requires integrating Equation 5.3-4 over $z_2, z_3, ..., z_{n-1}$, recalling that $z_{i-1} < z_i < 1$. The result is

$$\begin{aligned} f_{Z_1 Z_n}(z_1, z_n) &= n(n-1)(z_n - z_1)^{n-2} \quad \text{for } 0 < z_1 < z_n < 1, n \geq 2 \\ &= 0, \text{ else.} \end{aligned} \tag{5.3-7}$$

Consider now two new RVs $V_{1n} \triangleq Z_n - Z_1, W \triangleq Z_n$. To find the pdf of $f_{VW}(v, w)$, we consider the transformation $v = z_n - z_1, w = z_n; 0 < v < w < 1$. The Jacobian magnitude of this transformation is 1 and the only solution to this transformation is $z_1^{(r)} = w - v$; $z_n^{(r)} = w$. Hence

$$\begin{aligned} f_{V_{1n}W}(v, w) &= n(n-1)v^{n-2} \quad \text{for } 0 < w - v < w < 1, n \geq 2 \\ &= 0, \text{ else.} \end{aligned}$$

To get the pdf of V_{1n} alone, we integrate out with respect to w. To help with the integration, we note that the two inequalities $w - v > 0$ and $w < 1$ suggest the triangular region of integration shown in Figure 5.3-3

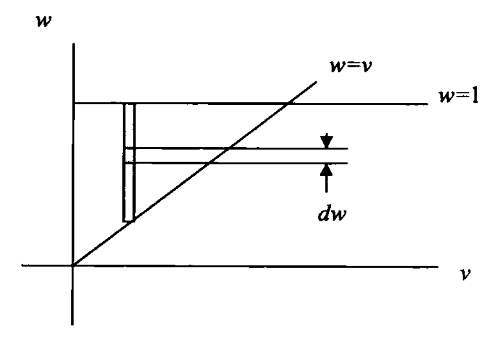


Figure 5.3-3 Region of integration for computing the probability density function of V_{1n}.



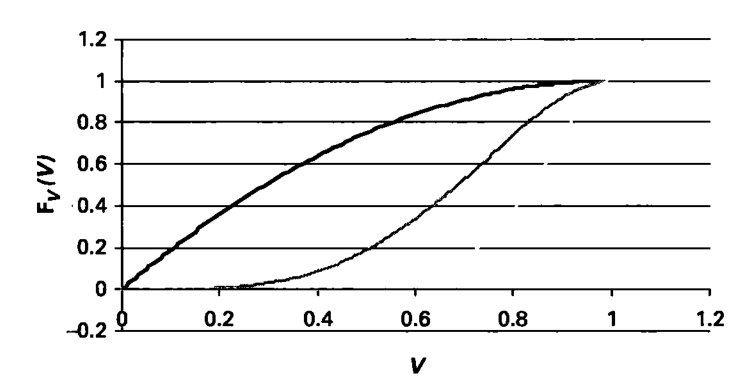


Figure 5.3-4 The beta CDF (Equation 5.3-9) for $n = 2$ (top curve); $n = 4$ (middle curve); $n = 10$ (bottom curve).

Thus starting with $f_{V_{1n}}(v) = n(n-1)v^{n-2}\int_v^1 dw$ we obtain

$$f_{V_{1n}}(v) = \begin{cases} n(n-1)v^{n-2}(1-v), & \text{for } 0 < v < 1, n \geq 2 \\ 0, & \text{else} \end{cases} \tag{5.3-8}$$

This pdf is a special case of the beta density given in Section 2.4 with $\alpha = n - 2, \beta = 1$. The distribution function is the probability that the area spread between the largest and the smallest is less than or equal to v. It is readily computed as

$$F_{V_{1n}}(v) = \begin{cases} nv^{n-1} - (n-1)v^n, & 0 < v < 1 \\ 1, & v > 1 \\ 0, & v < 0. \end{cases} \tag{5.3-9}$$

The beta CDF is shown in Figure 5.3-4 for various values of n.

Example 5.3-2

(*area between any ordered RVs*) We can extend the above results to computing the density of the areas under $f_X(x)$ between any ordered RVs, not necessarily between the first and the last. We generalize the notation slightly so that

$$V_{lm} \triangleq Z_m - Z_l = \int_{-\infty}^{Y_m} f_X(x)dx - \int_{-\infty}^{Y_l} f_X(x)dx = \int_{Y_l}^{Y_m} f_X(x)dx, m > l. \tag{5.3-10}$$

Consider $0 < Z_1 < Z_2 < Z_3 < 1$ with $f_{Z_1 Z_2 Z_3}(z_1, z_2, z_3) = 3!, 0 < z_1 < z_2 < z_3 < 1$. We first consider the density, $f_{V_{23}}(v)$, of the RV $V_{23} = Z_3 - Z_2$. Since this involves only Z_2 and Z_3, we must compute $f_{Z_2 Z_3}(z_2, z_3)$ from $f_{Z_1 Z_2 Z_3}(z_1, z_2, z_3)$. This is done as

$$f_{Z_2 Z_3}(z_2, z_3) = \begin{cases} 3!\int_0^{z_2} dz_1 = 3!z_2, & \text{for } 0 < z_2 < z_3 < 1 \\ 0, & \text{else} \end{cases}$$

To compute $f_{V_{23}}(v)$ from $f_{Z_2Z_3}(z_2, z_3)$, we define an auxiliary RV $B \triangleq Z_2$ with realizations β and an appropriate set of functional equations. In this case a suitable set of functional equations are $v \triangleq z_3 - z_2, \beta \triangleq z_2$ with roots $z_2^{(r)} = \beta, z_3^{(r)} = v + \beta$. The reader will recognize that $\beta \triangleq z_2$ serves as the auxiliary variable. Then, using the so-called direct formula yields

$$\begin{aligned} f_{BV_{23}}(\beta, v) &= f_{Z_2Z_3}(\beta, v+\beta)/|J| = 3!\beta \text{ for } 0 < \beta < 1 - v \\ &= 0, \text{ else} \end{aligned}$$

as the Jacobian magnitude $|J|$ of the transformation is unity.

Finally, integrating over the auxiliary variable β yields

$$\begin{aligned} f_{V_{23}}(v) &= 3! \int_0^{1-v} \beta d\beta = \tfrac{3!(1-v)^2}{2!}, 0 < v < 1 \\ &= 0, \text{ else.} \end{aligned} \tag{5.3-11}$$

To compute $f_{V_{12}}(v)$ from $f_{Z_1Z_2Z_3}(z_1, z_2, z_3)$, we proceed in the same fashion. Here we find that $f_{Z_1Z_2}(z_1, z_2)$ is given by $f_{Z_1Z_2}(z_1, z_2) = 3!(1 - z_2)$ for $0 < z_1 < z_2 < 1$ and 0 else. Then, using the transformation $v \triangleq z_2 - z_1, \beta \triangleq z_1$ we get the result

$$\begin{aligned} f_{V_{12}}(v) &= 3! \int_0^{1-v} (1 - v - \beta) d\beta = \tfrac{3!(1-v)^2}{2!}, 0 < v < 1 \\ &= 0, \text{ else.} \end{aligned} \tag{5.3-12}$$

We leave the details to the reader.

The general case is given by the following: let V_{lm} denote the probability area under $f_X(x)$ between Y_l and Y_m of the samples ordered by size $Y_1, Y_2, ..., Y_n$ drawn from the pdf $f_X(x)$. Then the pdf of V_{lm} is given by

$$\begin{aligned} f_{V_{lm}}(v) &= \tfrac{n!}{(m-l-1)!(n-m+l)!} v^{m-l-1}(1-v)^{n-m+l}, 0 < v < 1 \\ &= 0, \text{ else.} \end{aligned} \tag{5.3-13}$$

Example 5.3-3

(*expected value of area under $f_X(x)$ between ordered samples*) Consider the area RVs $0 < Z_1 < Z_2 < Z_3 < 1$, where Z_i is given in Equation 5.3-2. We wish to compute $E[Z_i]$ for $1 = 1, 2, 3$, expecting that $E[Z_1] < E[Z_2] < E[Z_3]$. We find that the marginal pdf's $f_{Z_i}(z_i), i = 1, 2, 3$, are computed as

$$\begin{aligned} f_{Z_1}(z_1) &= \int_{z_1}^1 \int_{z_1}^{z_3} f_{Z_1Z_2Z_3}(z_1, z_2, z_3) dz_2 dz_3 = \tfrac{3!}{2!}(1 - z_1)^2 \\ f_{Z_2}(z_2) &= \int_{z_2}^1 \int_0^{z_2} f_{Z_1Z_2Z_3}(z_1, z_2, z_3) dz_1 dz_3 = 3! z_2 (1 - z_2) \\ f_{Z_3}(z_3) &= \int_0^{z_3} \int_0^{z_2} f_{Z_1Z_2Z_3}(z_1, z_2, z_3) dz_1 dz_2 = \tfrac{3!}{2!} z_3^2. \end{aligned}$$

From these results it follows that

$$\begin{aligned} E[Z_1] &= \int_0^1 z f_{Z_1}(z) dz = \tfrac{1}{4} = \tfrac{1}{3+1} \\ E[Z_2] &= \int_0^1 z f_{Z_2}(z) dz = \tfrac{2}{4} = \tfrac{2}{3+1} \\ E[Z_3] &= \int_0^1 z f_{Z_3}(z) dz = \tfrac{3}{4} = \tfrac{3}{3+1}, \end{aligned}$$

which suggests that in the general case, that is $0 < Z_1 < Z_2 < \cdots < Z_n < 1$,

$$E[Z_i] = \frac{i}{n+1}. \tag{5.3-14}$$

The general case can be obtained by induction.

Example 5.3-4

(*moments of area between ordered samples*) Consider the area between two adjacent ordered samples. This is given by $V_{i,i+1} = Z_{i+1} - Z_i$. The pdf of $V_{i,i+1}$ is given by Equation 5.3-13 by letting $m = i+1, l = i$, which yields $f_{V_{i,i+1}}(v) = n(1-v)^{n-1}$, for $0 < v < 1$ and 0, else. Note that this result is independent of i. From this we compute

$$E[V_{i,i+1}] = n\int_0^1 v(1-v)^{n-1}dv = n\frac{\Gamma(2)\Gamma(n)}{\Gamma(n+2)} = \frac{1}{n+1},$$

where the gamma function $\Gamma(j) = (j-1)!$ for $j = 1, 2, \ldots$ and use was made of tables of integrals (see for example formula 497, p.67 in *A Short Table of Integrals* by B. O. Peirce and R. M. Foster, Ginn and Company, New York, 1956). The integral can also be found online at several places, including www.wolframalpha.com (type 'integral' at the prompt). Likewise

$$E[V_{i,i+1}^2] = n\int_0^1 v^2(1-v)^{n-1}dv = n\frac{\Gamma(3)\Gamma(n)}{\Gamma(n+3)} = n\frac{2!(n-1)!}{(n+2)!} = \frac{2}{(n+2)(n+1)}.$$

Hence $\sigma^2_{V_{i,i+1}} = \frac{2}{(n+1)(n+2)} - \frac{1}{(n+1)^2} \approx \frac{1}{(n+1)^2}$ for $n >> 1$.

To compute the variances $\sigma^2_{Z_i}, i = 1, ..., n$, we first compute $E[Z_i^2]$ for $i = 1, ..., n$ as

$$\begin{aligned}
&E[Z_1^2] = n!\int_0^1\int_0^{z_n}\int_0^{z_{n-1}}\cdots(\int_0^{z_2} z_1^2 dz_1)\cdots dz_{n-1}dz_n = 2\left((n+2)(n+1)\right)^{-1}\\
&E[Z_2^2] = n!\int_0^1\int_0^{z_n}\int_0^{z_{n-1}}\cdots\int_0^{z_3} z_2^2\int_0^{z_2} dz_1dz_2\cdots dz_n = 6\left((n+2)(n+1)\right)^{-1}\\
&\vdots\\
&E[Z_n^2] = n!\int_0^1 z_n^2\int_0^{z_n}\int_0^{z_{n-1}}\cdots\int_0^{z_3}\int_0^{z_2} dz_1dz_2\cdots dz_n\\
&= n(n+1)/\left((n+2)(n+1)\right)^{-1}.
\end{aligned}$$

It follows that

$$E[Z_i^2] = \frac{i(i+1)}{(n+1)(n+2)} \text{ for } i = 1, ..., n$$

and the variances, computed as $\sigma_i^2 = E[Z_i^2] - E^2[Z_i]$, yield

$$\sigma^2_{Z_i} = \frac{i(i+1)}{(n+1)(n+2)} - \frac{i^2}{(n+1)^2} \approx \frac{i}{(n+1)^2} \text{ for } n >> 1.$$

Thus for large n, $\sigma^2_{Z_i} \approx E[Z_i]/(n+1)$.

Example 5.3-5

(*Estimating range of boot sizes*) Military boots need to be ordered for fresh Army recruits but the manufacture needs to know what range of boot sizes will be required. It is suggested that a random sample of n recruits be measured for required boot size. What is the minimum value of n that will cover at least 95 percent of the boot-size needs of the recruits?

Solution Let $\{X_i, i = 1, \ldots, n\}$ denote the i.i.d. set of boot sizes of the n recruits drawn from a population with (unknown) pdf $f_X(x)$ and let $\{Y_i, i = 1, \ldots, n\}$ denote the order statistics of the observations. With $V_{1n} = \int_{Y_1}^{Y_n} f_X(x)dx$ we need to solve $P[V_{1n} \geq 0.95] = \delta$, where δ is a measure of the reliability of our estimate of n, that is, in 100δ percent of the time, the number n will indeed be the minimum sample size required for estimating the boot needs of the recruits. Using $P[V_{1n} \leq 0.95] = 1 - \delta$ and Equation 5.3-9 we compute $n = 93$ for $\delta = 0.95$ and $n = 114$ for $\delta = 0.98$. The solution is obtained numerically using Excel™. Note that the result is independent of the size of the recruit army.

5.4 EXPECTATION VECTORS AND COVARIANCE MATRICES†

Definition 5.4-1 The expected value of the (column) vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ is a vector $\boldsymbol{\mu}$ (or $\overline{\mathbf{X}}$) whose elements $\mu_1, \ldots, \mu_n$ are given by

$$\mu_i \triangleq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i f_{\mathbf{X}}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n. \tag{5.4-1}$$

Equivalently with

$$f_{X_i}(x_i) \triangleq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})\, dx_1 \ldots dx_{i-1}\, dx_{i+1} \ldots dx_n$$

the marginal pdf of X_i, we can write

$$\mu_i = \int_{-\infty}^{\infty} x_i f_{X_i}(x_i)\, dx_i \qquad i = 1, \ldots, n. \quad \blacksquare$$

Definition 5.4-2 The covariance matrix $\mathbf{K}$ associated with a real random vector $\mathbf{X}$ is the expected value of the outer vector product $(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T$, that is‡,

$$\mathbf{K} \triangleq E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]. \tag{5.4-2}$$

We have for the (i, j)th component

$$K_{ij} \triangleq E[(X_i - \mu_i)(X_j - \mu_j)]$$

†This section requires some familiarity with matrix theory.

‡We temporarily dispense with adding identifying subscripts on the mean, covariance and other vector parameters since it is clear we are dealing only with the RV $\mathbf{X}$.

$$= E[(X_j - \mu_j)(X_i - \mu_i)]$$
$$= K_{ji} \qquad i, j = 1, \ldots, n. \tag{5.4-3}$$

In particular with $\sigma_i^2 \triangleq K_{ii}$, we can write $\mathbf{K}$ in expanded form as

$$\mathbf{K} = \begin{bmatrix} \sigma_1^2 & & \cdots & & K_{1n} \\ & \ddots & & & \\ \vdots & & \sigma_i^2 & & \vdots \\ & & & \ddots & \\ K_{n1} & & \cdots & & \sigma_n^2 \end{bmatrix}. \ \blacksquare \tag{5.4-4}$$

If $\mathbf{X}$ is real, all the elements of $\mathbf{K}$ are real. Also since $K_{ij} = K_{ji}$, real-valued covariance matrices fall within the class of matrices called *real symmetric* (r.s.). Such matrices fall within the larger class of Hermitian matrices.† Real symmetric matrices have many interesting properties, several of which we shall discuss in the next section.

The diagonal elements σ_i^2 are the variances associated with the individual RVs X_i for $i = 1, \ldots, n$. The covariance matrix $\mathbf{K}$ is closely related to the correlation matrix $\mathbf{R}$ defined by

$$\mathbf{R} \triangleq E[\mathbf{X}\mathbf{X}^T]. \tag{5.4-5}$$

Indeed expanding Equation 5.4-2 yields

$$\mathbf{K} = \mathbf{R} - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

or

$$\mathbf{R} = \mathbf{K} + \boldsymbol{\mu}\boldsymbol{\mu}^T. \tag{5.4-6}$$

The correlation matrix $\mathbf{R}$ is also real symmetric for a real-valued random vector and is sometimes called the *autocorrelation* matrix. Random vectors are often classified according to whether they are uncorrelated, orthogonal, or independent.

Definition 5.4-3 Consider two real n-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ with respective mean vectors $\boldsymbol{\mu}_\mathbf{X}$ and $\boldsymbol{\mu}_\mathbf{Y}$. Then if the expected value of their *outer product* satisfies

$$E\{\mathbf{X}\mathbf{Y}^T\} = \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}_\mathbf{Y}{}^T, \tag{5.4-7}$$

$\mathbf{X}$ and $\mathbf{Y}$ are said to be *uncorrelated.* If

$$E\{\mathbf{X}\mathbf{Y}^T\} = \mathbf{0} \quad \text{(an } n \times n \text{ matrix of all zeros)}, \tag{5.4-8}$$

$\mathbf{X}$ and $\mathbf{Y}$ are said to be *orthogonal.*

†The class of $n \times n$ matrices for which $K_{ij} = K_{ji}^*$. For a thorough discussion of the properties of such matrices see [5-2]. When $\mathbf{X}$ is complex, the covariance is generally not r.s. but is Hermitian.

Note that in the orthogonal case $E\{X_iY_j\} = 0$ for all $0 \le i,\ j \le n$. Thus, the expected value of the *inner product* is zero, that is, $\mathbf{E}[\mathbf{X}^T\mathbf{Y}] = 0$, which reminds us of the meaning of orthogonality for two ordinary (nonrandom) vectors, that is, $\mathbf{x}^T\mathbf{y} = 0$.

Finally if

$$f_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}}(\mathbf{y}), \tag{5.4-9}$$

$\mathbf{X}$ and $\mathbf{Y}$ are said to be *independent*. ■

Independence always implies uncorrelatedness but the converse is not generally true. An exception is the multidimensional Gaussian pdf to be presented in Section 5.6. It is often difficult, in practice, to show that two random vectors are independent. However, statistical tests exist to determine, within prescribed confidence levels, the extent to which they are correlated.

Example 5.4-1

(*almost independent RVs*) Consider two RVs X_1 and X_2 with joint pdf $f_{X_1X_2}(x_1, x_1) = x_1 + x_2$ for $0 < x_1 \le 1$, $0 < x_2 \le 1$, and zero elsewhere. We find that while X_1 and X_2 are not independent, they are essentially uncorrelated. To demonstrate this, we shall compute $E[(X_1 - \mu_1)(X_2 - \mu_2)]$ as

$$K_{12} = K_{21} = R_{21} - \mu_2\mu_1.$$

We first compute

$$\mu_1 = \mu_2 = \iint_S x(x+y)\,dx\,dy = 0.583,$$

where $S = \{(x_1, x_2)\colon 0 < x_1 \le 1,\ 0 < x_2 \le 1\}$.

Next we compute the correlation products

$$R_{12} = R_{21} \triangleq \iint_S xy(x+y)\,dx\,dy = 0.333.$$

Hence $K_{12} = K_{21} = 0.333 - (0.583)^2 = -0.007$. Also we compute

$$\sigma_1^2 = \int_0^1 x^2(x + \tfrac{1}{2})\,dx - (0.583)^2 = 0.4167 - 0.34 = 0.077.$$

Hence the correlation coefficient (normalized covariance) is computed to be $\rho = K_{12}/\sigma_1\sigma_2 = -0.091$. For the purpose of predicting X_2 by observing X_1, or vice versa, one may consider these RVs as being uncorrelated. Indeed the prediction error ε in Equation 4.3-22 from Example 4.3-4 is 0.076. Were X_1, X_2 truly uncorrelated, the prediction error would have been 0.077. The covariance matrix $\mathbf{K}$ for this case is

$$\mathbf{K} = \begin{bmatrix} 0.077 & -0.007 \\ -0.007 & 0.077 \end{bmatrix} = 0.077 \begin{bmatrix} 1 & -0.09 \\ -0.09 & 1 \end{bmatrix}.$$

5.5 PROPERTIES OF COVARIANCE MATRICES

Since covariance matrices are r.s., we study some of the properties of such matrices. Let $\mathbf{M}$ be any $n \times n$ r.s. matrix. The quadratic form associated with $\mathbf{M}$ is the scalar $q(\mathbf{z})$ defined by

$$q(\mathbf{z}) \triangleq \mathbf{z}^T\mathbf{M}\mathbf{z}, \tag{5.5-1}$$

where $\mathbf{z}$ is any column vector. A matrix $\mathbf{M}$ is said to be *positive semidefinite* (p.s.d.) if

$$\mathbf{z}^T\mathbf{M}\mathbf{z} \geq 0$$

for all $\mathbf{z}$. If the inequality is strict, i.e. $\mathbf{z}^T\mathbf{M}\mathbf{z} > 0$ for all $\mathbf{z} \neq \mathbf{0}$, $\mathbf{M}$ is said to be *positive definite* (p.d.). A covariance matrix $\mathbf{K}$ is always (at least) p.s.d. since for any vector $\mathbf{z} \triangleq (z_1, \ldots, z_n)^T$

$$\begin{aligned} 0 &\leq E\{[\mathbf{z}^T(\mathbf{X} - \boldsymbol{\mu})]^2\} \\ &= \mathbf{z}^T E[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]\mathbf{z} \\ &= \mathbf{z}^T\mathbf{K}\mathbf{z} \end{aligned} \tag{5.5-2}$$

We shall show later that when $\mathbf{K}$ is full-rank, then $\mathbf{K}$ is p.d.

We now state some definitions and theorems (most without proof) from linear algebra [5-2, Chapter 4] that we shall need for developing useful operations on covariance matrices.

Definition 5.5-1 The *eigenvalues* of an $n \times n$ matrix $\mathbf{M}$ are those numbers λ for which the characteristic equation $\mathbf{M}\boldsymbol{\phi} = \lambda\boldsymbol{\phi}$ has a solution $\boldsymbol{\phi} \neq \mathbf{0}$. The column vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)^T$ is called an *eigenvector*.

Eigenvectors are often normalized so that $\boldsymbol{\phi}^T\boldsymbol{\phi} \triangleq ||\boldsymbol{\phi}||^2 = 1$. ■

Theorem 5.5-1 The number λ is an eigenvalue of the square matrix $\mathbf{M}$ if and only if $\det(\mathbf{M} - \lambda\mathbf{I}) = 0$.† ■

Example 5.5-1

(*eigenvalues*) Consider the matrix

$$\mathbf{M} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}.$$

The eigenvalues are obtained with the help of Theorem 5.5-1, that is,

$$\det \begin{bmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix} = (4-\lambda)^2 - 4 = 0,$$

whence

$$\lambda_1 = 6, \qquad \lambda_2 = 2.$$

†det is short for determinant and $\mathbf{I}$ is the identity matrix.

The (normalized) eigenvector associated with $\lambda_1 = 6$ is obtained from

$$(\mathbf{M} - 6\mathbf{I})\boldsymbol{\phi} = 0,$$

which, written out as a system of equations, yields

$$\left.\begin{matrix} -2\phi_1 + 2\phi_2 = 0 \\ -2\phi_1 + 2\phi_2 = 0 \end{matrix}\right\} \Rightarrow \boldsymbol{\phi}_1 = \frac{1}{\sqrt{2}}(1,1)^T.$$

The double arrow $\Rightarrow$ means "implies that." The eigenvector associated with $\lambda_2 = 2$, following the same procedure as above, is found from

$$\left.\begin{matrix} 2\phi_1 + 2\phi_2 = 0 \\ 2\phi_1 + 2\phi_2 = 0 \end{matrix}\right\} \Rightarrow \boldsymbol{\phi}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Not all $n \times n$ matrices have n distinct eigenvalues or n eigenvectors. Sometimes a matrix can have fewer than n distinct eigenvalues but still have n distinct eigenvectors.

Definition 5.5-2 Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are called *similar* if there exists an $n \times n$ invertible matrix $\mathbf{T}$, i.e. det $\mathbf{T} \neq 0$, such that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{B}. \quad \blacksquare \tag{5.5-3}$$

Theorem 5.5-2 An $n \times n$ matrix $\mathbf{M}$ is similar to a diagonal matrix if and only if $\mathbf{M}$ has n linearly independent eigenvectors. ■

Theorem 5.5-3 Let $\mathbf{M}$ be an r.s. matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $\mathbf{M}$ has n mutually orthogonal unit eigenvectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$[†]. ■

Discussion. Since $\mathbf{M}$ has n mutually orthogonal (and therefore independent) unit eigenvectors, it is similar to some diagonal matrix $\boldsymbol{\Lambda}$ under a suitable transformation $\mathbf{T}$. What are $\boldsymbol{\Lambda}$ and $\mathbf{T}$? The answer is furnished by the following important theorem.

Theorem 5.5-4 Let $\mathbf{M}$ be a real symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then $\mathbf{M}$ is similar to the diagonal matrix $\boldsymbol{\Lambda}$ given by

$$\boldsymbol{\Lambda} \triangleq \begin{bmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{bmatrix}$$

under the transformation

$$\mathbf{U}^{-1}\mathbf{M}\mathbf{U} = \boldsymbol{\Lambda}, \tag{5.5-4}$$

where $\mathbf{U}$ is a matrix whose columns are the corresponding[‡] orthogonal unit eigenvectors $\boldsymbol{\phi}_i$, $i = 1, \ldots, n$, of $\mathbf{M}$. Thus,

$$\mathbf{U} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n). \tag{5.5-5}$$

Moreover, it can be shown that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ (and that $\mathbf{U}^T = \mathbf{U}^{-1}$) so that Equation 5.5-4 can be written as

$$\mathbf{U}^T\mathbf{M}\mathbf{U} = \boldsymbol{\Lambda}. \quad \blacksquare \tag{5.5-6}$$

[†] Orthogonal eigenvectors ϕ_i such that $||\phi_i|| = 1$ are said to be orthonormal.

[‡] That is, $\boldsymbol{\phi}_i$ goes with λ_i for $i = 1, \ldots, n$.

Discussion. Matrices such as $\mathbf{M}$, which satisfy $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, are called *unitary*. They have the property of *distance preservation* in the following sense: Consider a vector $\mathbf{x} = (x_1, \ldots, x_n)^T$. The Euclidean distance of $\mathbf{x}$ from the origin is

$$||\mathbf{x}|| \triangleq (\mathbf{x}^T\mathbf{x})^{1/2},$$

where $||\mathbf{x}||$ is called the *norm* of $\mathbf{x}$. Now consider the transformation $\mathbf{y} = \mathbf{U}\mathbf{x}$, where $\mathbf{U}$ is unitary. Then

$$||\mathbf{y}||^2 = \mathbf{y}^T\mathbf{y} = \mathbf{x}^T\mathbf{U}^T\mathbf{U}\mathbf{x} = ||\mathbf{x}||^2.$$

Thus, the new vector $\mathbf{y}$ has the same distance from the origin as the old vector $\mathbf{x}$ under the transformation $\mathbf{y} = \mathbf{U}\mathbf{x}$.

Since a covariance matrix $\mathbf{K}$ of a real random vector is real symmetric, it can be readily diagonalized according to Equation 5.5-6 once $\mathbf{U}$ is known. The columns of $\mathbf{U}$ are just the normalized eigenvectors of $\mathbf{K}$ and these can be obtained once the eigenvalues are known. The diagonalization of covariance matrices is a very important procedure in applied probability theory. It is used to transform correlated RVs into uncorrelated RVs and, in the Normal case, it transforms correlated RVs into independent RVs.

Example 5.5-2

(*decorrelation of random vectors*) A random vector $\mathbf{X} = (X_1, X_2, X_3)^T$ has covariance matrix†

$$\mathbf{K}_{\mathbf{XX}} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}.$$

Design an invertible linear transformation that will generate from $\mathbf{X}$ a new random vector $\mathbf{Y}$ whose components are uncorrelated.

Solution First we compute the eigenvalues by solving the equation $\det(\mathbf{K}_{\mathbf{XX}} - \lambda\mathbf{I}) = 0$. This yields $\lambda_1 = 2$, $\lambda_2 = 2 + \sqrt{2}$, $\lambda_3 = 2 - \sqrt{2}$. Next we compute the three orthogonal eigenvectors by solving the equation $(\mathbf{K}_{\mathbf{XX}} - \lambda_i\mathbf{I})\boldsymbol{\phi}_i = \mathbf{0}$, $i = 1, 2, 3$ and normalize these to create eigenvectors of unit norm. Unit normalization is achieved by dividing each component of the eigenvector by the norm of the eigenvector. This yields

$$\boldsymbol{\phi}_1 = \left(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T,$$

$$\boldsymbol{\phi}_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{2}, \frac{1}{2}\right)^T,$$

$$\boldsymbol{\phi}_3 = \left(\frac{1}{\sqrt{2}}, \frac{1}{2}, -\frac{1}{2}\right)^T.$$

†Here we add subscripts to $\mathbf{K}$ to help distinguish the covariance matrix of one random variable from that of another.

Now we create the eigenvector matrix $\mathbf{U} = [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2\ \boldsymbol{\phi}_3]$ that, upon transposing, becomes an appropriate transformer to make the components of $\mathbf{Y}$ uncorrelated. With

$$\mathbf{A} = \mathbf{U}^T = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix},$$

the transformation $\mathbf{Y} = \mathbf{AX}$ yields the components

$$Y_1 = \frac{1}{\sqrt{2}}(X_2 + X_3)$$

$$Y_2 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{2}X_2 + \frac{1}{2}X_3$$

$$Y_3 = \frac{1}{\sqrt{2}}X_1 + \frac{1}{2}X_2 - \frac{1}{2}X_3.$$

The covariance of $\mathbf{Y}$ is given by

$$\mathbf{K}_{\mathbf{YY}} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2+\sqrt{2} & 0 \\ 0 & 0 & 2-\sqrt{2} \end{bmatrix}.$$

Actually we could go one step further; by scaling the three components of $\mathbf{Y}$, separately, we can make the variance (average AC power) the same in each scaled component. This process is called *whitening* and is discussed in greater detail below. Clearly if Y_1 is scaled proportional to $\frac{1}{\sqrt{\lambda_1}}$, Y_2 is scaled proportional to $\frac{1}{\sqrt{\lambda_2}}$, and Y_3 is scaled proportional to $\frac{1}{\sqrt{\lambda_3}}$, all three outputs will have the same power.

If $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$ are the orthogonal unit eigenvectors of a real symmetric matrix $\mathbf{M}$, then the system of equations

$$\mathbf{M}\boldsymbol{\phi}_1 = \lambda_1\boldsymbol{\phi}_1$$

$$\vdots$$

$$\mathbf{M}\boldsymbol{\phi}_n = \lambda_n\boldsymbol{\phi}_n$$

can be compactly written as

$$\mathbf{MU} = \mathbf{U}\mathbf{\Lambda}. \tag{5.5-7}$$

The next theorem establishes a relation between the eigenvalues of an r.s. matrix and its positive definite character.

Theorem 5.5-5 A real symmetric matrix $\mathbf{M}$ is positive definite if and only if all its eigenvalues are positive.

Proof First let $\lambda_i > 0$, $i = 1, \ldots, n$. Then with the linear transformation $\mathbf{x} \triangleq \mathbf{U}\mathbf{y}$ we can write for any vector $\mathbf{x}$

$$\begin{aligned}\mathbf{x}^T\mathbf{M}\mathbf{x} &= (\mathbf{U}\mathbf{y})^T\mathbf{M}(\mathbf{U}\mathbf{y}) \\ &= \mathbf{y}^T\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{y} \\ &= \mathbf{y}^T\mathbf{\Lambda}\mathbf{y} \\ &= \sum_{i=1}^{n}\lambda_i y_i^2 > 0 \end{aligned} \tag{5.5-8}$$

unless $\mathbf{y} = 0$. But if $\mathbf{y} = 0$, then from $\mathbf{x} = \mathbf{U}\mathbf{y}$, $\mathbf{x} = 0$ as well. Hence we have shown that $\mathbf{M}$ is p.d. if $\lambda_i > 0$ for all i. Conversely, we must show that if $\mathbf{M}$ is p.d., then all $\lambda_i > 0$. Thus, for any $\mathbf{x} \neq 0$

$$0 < \mathbf{x}^T\mathbf{M}\mathbf{x}. \tag{5.5-9}$$

In particular, Equation 5.5-9 must hold for $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$. But

$$0 < \boldsymbol{\phi}_i^T\mathbf{M}\boldsymbol{\phi}_i = \lambda_i, \qquad i = 1, \ldots, n.$$

Hence $\lambda_i > 0$, $i = 1, \ldots, n$. Thus, a p.d. covariance matrix $\mathbf{K}$ will have all positive eigenvalues. Also since its determinant $\det(\mathbf{K})$ is the product of its eigenvalues, $\det(\mathbf{K}) > 0$. Thus when $\mathbf{K}$ is full-rank, it is p.d. ■

Whitening Transformation

We are given a zero-mean $n \times 1$ random vector $\mathbf{X}$ with positive definite covariance matrix $\mathbf{K_{XX}}$ and wish to find a transformation $\mathbf{Y} = \mathbf{CX}$ such that $\mathbf{K_{YY}} = \mathbf{I}$. The matrix $\mathbf{C}$ is called a *whitening transform* and process of going from $\mathbf{X}$ to $\mathbf{Y}$ is called a *whitening transformation.* Let the n unit eigenvectors and eigenvalues of $\mathbf{K_{XX}}$ be denoted, respectively, by $\boldsymbol{\phi}_i, \lambda_i, i = 1, \cdots, n$. Then the characteristic equation $\mathbf{K_{XX}}\boldsymbol{\phi}_i = \lambda_i\boldsymbol{\phi}_i, i = 1, \cdots, n$ can be compactly written as $\mathbf{K_{XX}U} = \mathbf{U\Lambda}$, where $\mathbf{U} \triangleq [\boldsymbol{\phi}_1\ \boldsymbol{\phi}_2 \cdots \boldsymbol{\phi}_n]$ and $\mathbf{\Lambda} \triangleq \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$. Since $\mathbf{K_{XX}}$ is p.d., all its eigenvalues are positive and the matrix $\Lambda^{-1/2} \triangleq \operatorname{diag}(1/\sqrt{\lambda_1}, 1/\sqrt{\lambda_2}, \cdots, 1/\sqrt{\lambda_n})$ exists and is well defined. Now consider the transformation $\mathbf{Y} = \mathbf{CX} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{X}$. Then

$\mathbf{K_{YY}} = E[\mathbf{YY}^T] = E[\mathbf{CXX}^T\mathbf{C}^T] = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T E[\mathbf{XX}^T]\mathbf{U}\Lambda^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T\mathbf{K_{XX}}\ \mathbf{U}\Lambda^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{K_{XX}U})\Lambda^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{U}\Lambda)\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}(\mathbf{U}^T\mathbf{U})\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} = \mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}\mathbf{\Lambda}^{-1/2} = \mathbf{I}$, since $\mathbf{U}^T\mathbf{U} = \mathbf{I}$.

Example 5.5-3

(whitening transformation) In Example 5.5-2 we considered the random vector $\mathbf{X}$ with covariance matrix and eigenvector matrices, respectively

$$\mathbf{K_{XX}} = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/2 & 1/2 \\ 1/\sqrt{2} & 1/2 & -1/2 \end{bmatrix} = \mathbf{U}^T$$

with

$$\mathbf{\Lambda} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2+\sqrt{2} & 0 \\ 0 & 0 & 2-\sqrt{2} \end{bmatrix}.$$

Then

$$\mathbf{Y} = \begin{bmatrix} 1/\sqrt{2} & 0 & 0 \\ 0 & (2+\sqrt{2})^{-1/2} & 0 \\ 0 & 0 & (2-\sqrt{2})^{-1/2} \end{bmatrix} \begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1/-2 & 1/2 \\ 1/\sqrt{2} & 1/2 & -1/2 \end{bmatrix} \mathbf{X}$$

is the appropriate whitening transformation. Whitening transformations are especially useful in the simultaneous diagonalization of two covariance matrices.†

5.6 THE MULTIDIMENSIONAL GAUSSIAN (NORMAL) LAW

The general n-dimensional Gaussian law has a rather forbidding mathematical appearance upon first acquaintance but is, fortunately, rather easily seen as an extension of the one-dimensional Gaussian pdf. Indeed we already introduced the two-dimensional Gaussian pdf in Section 4.3 but there we did not infer it from the general case. Here we consider the general case from which we shall be able to infer all special cases. We already know that if X is a (scalar) Gaussian RV with mean μ and variance σ^2, its pdf is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

First, we consider a random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ with *independent* components X_i, $i = 1, \ldots, n$, each distributed as $N(\mu_i, \sigma_i{}^2)$. Then the pdf of $\mathbf{X}$ is the product of the individual pdf's of $X_1, \ldots, X_n$, that is,

$$\begin{aligned} f_{\mathbf{X}}(x_1, \ldots, x_n) &= \prod_{i=1}^{n} f_{X_i}(x_i) \\ &= \frac{1}{(2\pi)^{n/2}\sigma_1 \ldots \sigma_n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right], \end{aligned} \tag{5.6-1}$$

†Such diagonalizations occur in a branch of applied probability called pattern recognition. In particular, if one is trying to distinguish between two classes of data, it is easier to do so when the data are represented by diagonal covariance matrices.

where μ_i, σ_i^2 are the mean and variance, respectively, of $X_i, i = 1, \ldots, n$. Equation 5.6-1 can be written compactly as

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}[\det(\mathbf{K_{XX}})]^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K}_{\mathbf{XX}}^{-1}(\mathbf{x}-\boldsymbol{\mu})], \tag{5.6-2}$$

where

$$\mathbf{K_{XX}} \triangleq \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}, \tag{5.6-3}$$

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, and $\det(\mathbf{K_{XX}}) = \prod_{i=1}^n \sigma_i^2$. Note that $\mathbf{K}_{\mathbf{XX}}^{-1}$ is merely

$$\mathbf{K}_{\mathbf{XX}}^{-1} = \begin{bmatrix} \sigma_1^{-2} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-2} \end{bmatrix}.$$

Note that because the X_i, $i = 1, \ldots, n$ are independent, the covariance matrix $\mathbf{K_{XX}}$ is diagonal, since

$$E[(X_i - \mu_i)^2] \triangleq \sigma_i^2 \quad i = 1, \ldots, n. \tag{5.6-4}$$

$$E[(X_i - \mu_i)(X_j - \mu_j)] = 0 \quad i \neq j. \tag{5.6-5}$$

Next we ask, what happens if $\mathbf{K_{XX}}$ is a positive definite covariance matrix that is not necessarily diagonal? Does Equation 5.6-2 with *arbitrary* p.d. covariance $\mathbf{K_{XX}}$ still obey the requirements of a pdf? If it does, we shall call $\mathbf{X}$ a *Normal* random vector and $f_{\mathbf{X}}(\mathbf{x})$ the *multidimensional Normal* pdf. To show that $f_{\mathbf{X}}(\mathbf{x})$ is indeed a pdf, we must show that

$$f_{\mathbf{X}}(\mathbf{x}) \geq 0 \tag{5.6-6a}$$

and

$$\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = 1 \tag{5.6-6b}$$

(We use the vector notation $d\mathbf{x} \triangleq dx_1\,dx_2 \ldots dx_n$ for a volume element.) We assume as always that $\mathbf{X}$ is real; that is, $X_1, \ldots, X_n$ are real RVs. To show that Equation 5.6-2 with arbitrary p.d. covariance matrix $\mathbf{K_{XX}}$ satisfies Equation 5.6-6a is simple and left as an exercise; to prove Equation 5.6-6b is more difficult, and follows here.

Proof of Equation 5.6-6b when $f_{\mathbf{X}}(\mathbf{x})$ is as in Equation 5.6-2 and $\mathbf{K_{XX}}$ is an arbitrary p.d. covariance matrix. We note that with $\mathbf{z} \triangleq \mathbf{x} - \boldsymbol{\mu}$, Equation 5.6-2 can be written as

$$f_{\mathbf{X}}(\mathbf{x}) \triangleq \frac{1}{(2\pi)^{n/2}[\det(\mathbf{K_{XX}})]^{1/2}} \phi(\mathbf{z}),$$

where

$$\phi(\mathbf{z}) \triangleq \exp(-\tfrac{1}{2}\mathbf{z}^T\mathbf{K}_{\mathbf{XX}}^{-1}\mathbf{z}). \tag{5.6-7a}$$

With

$$\alpha \triangleq \int_{-\infty}^{\infty} \phi(\mathbf{z})d\mathbf{z}, \tag{5.6-7b}$$

we see that

$$\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \frac{\alpha}{(2\pi)^{n/2}[\det(\mathbf{K}_{\mathbf{XX}})]^{1/2}}.$$

Hence we need only evaluate α to prove (or disprove) Equation 5.6-6b.

From the discussion on *whitening transformation* we know that there exists an $n \times n$ matrix $\mathbf{C}$ such that $\mathbf{K}_{\mathbf{XX}} = \mathbf{C}\mathbf{C}^T$ and $\mathbf{C}^T\mathbf{K}_{\mathbf{XX}}^{-1}\mathbf{C} = \mathbf{I}$ (the identity matrix). Now consider the linear transformation

$$\mathbf{z} = \mathbf{C}\mathbf{y} \tag{5.6-8}$$

for use in Equation 5.6-7a. To understand the effect of this transformation, we note first that

$$\mathbf{z}^T\mathbf{K}_{\mathbf{XX}}^{-1}\mathbf{z} = \mathbf{y}^T\mathbf{C}^T\mathbf{K}_{\mathbf{XX}}^{-1}\mathbf{C}\mathbf{y} = ||\mathbf{y}||^2 = \sum_{i=1}^{n} y_i^2$$

so that $\phi(\mathbf{z})$ is given by

$$\phi(\mathbf{z}) = \prod_{i=1}^{n} \exp[-\tfrac{1}{2}y_i^2].$$

Next we use a result from advanced calculus (see Kenneth Miller, [5-5, p. 16]) that for a linear transformation such as in Equation 5.6-8 volume elements are related as

$$d\mathbf{z} = |\det(\mathbf{C})|d\mathbf{y},$$

where $d\mathbf{z} \triangleq dz_1 \dots dz_n$ and $d\mathbf{y} = dy_1 \dots dy_n$. Hence Equation 5.6-7b is transformed to

$$\begin{aligned}\alpha &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right) dy_1 \dots dy_n |\det(\mathbf{C})| \\ &= \left[\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right]^n |\det(\mathbf{C})| \\ &= [2\pi]^{n/2}|\det(\mathbf{C})|.\end{aligned}$$

But since $\mathbf{K}_{\mathbf{XX}} = \mathbf{C}\mathbf{C}^T$, $\det(\mathbf{K}_{\mathbf{XX}}) = \det(\mathbf{C})\det(\mathbf{C}^T) = [\det(\mathbf{C})]^2$ or

$$|\det(\mathbf{C})| = |\det(\mathbf{K}_{\mathbf{XX}})|^{1/2} = (\det(\mathbf{K}_{\mathbf{XX}}))^{1/2}.$$

Hence

$$\alpha = (2\pi)^{n/2}[\det(\mathbf{K}_{\mathbf{XX}})]^{1/2}$$

and

$$\frac{\alpha}{[2\pi]^{n/2}[\det(\mathbf{K_{XX}})]^{1/2}} = 1,$$

which proves Equation 5.6-6b. ■

Having established that

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}[\det(\mathbf{K_{XX}})]^{1/2}} \exp(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K}_{\mathbf{XX}}^{-1}(\mathbf{x}-\boldsymbol{\mu})) \tag{5.6-9}$$

indeed satisfies the requirements of a pdf and is a generalization of the univariate Normal pdf, we now ask what is the pdf of the random vector $\mathbf{Y}$ given by

$$\mathbf{Y} \triangleq \mathbf{AX}, \tag{5.6-10}$$

where $\mathbf{A}$ is a nonsingular $n \times n$ transformation. The answer is furnished by the following theorem.

Theorem 5.6-1 Let $\mathbf{X}$ be an n-dimensional Normal random vector with positive definite covariance matrix $\mathbf{K_{XX}}$ and mean vector $\boldsymbol{\mu}$. Let $\mathbf{A}$ be a nonsingular linear transformation in n dimensions. Then $\mathbf{Y} \triangleq \mathbf{AX}$ is an n-dimensional Normal random vector with covariance matrix $\mathbf{K_{YY}} \triangleq \mathbf{AK_{XX}A}^T$ and mean vector $\boldsymbol{\beta} \triangleq \mathbf{A}\boldsymbol{\mu}$. ■

Proof We use Equation 5.2-11, that is,

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i=1}^{r} \frac{f_{\mathbf{X}}(\mathbf{x}_i)}{|J_i|}, \tag{5.6-11}$$

where $\mathbf{Y}$ is some function of $\mathbf{X}$, that is, $\mathbf{Y} = \mathbf{g}(\mathbf{X}) \triangleq (g_1(\mathbf{X}), \ldots, g_n(\mathbf{X}))^T$, the $\mathbf{x}_i$, $i = 1, \ldots, r$, are the roots of the equation $\mathbf{g}(\mathbf{x}_i) - \mathbf{y} = \mathbf{0}$, and J_i is the Jacobian evaluated at the ith root, that is,

$$J_i = \det\left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right)\bigg|_{\mathbf{x}=\mathbf{x}_i} = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{vmatrix}_{\mathbf{x}=\mathbf{x}_i}. \tag{5.6-12}$$

Since we are dealing with a nonsingular *linear* transformation, the only solution to

$$\mathbf{Ax} - \mathbf{y} = \mathbf{0} \quad \text{is} \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{y}. \tag{5.6-13}$$

Also

$$J_i = \det\left(\frac{\partial(\mathbf{Ax})}{\partial \mathbf{x}}\right) = \det(\mathbf{A}). \tag{5.6-14}$$

Hence

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}[\det(\mathbf{K_{XX}})]^{1/2}|\det(\mathbf{A})|}\exp(-\tfrac{1}{2}(\mathbf{A}^{-1}\mathbf{y}-\boldsymbol{\mu})^T\mathbf{K}_{\mathbf{XX}}^{-1}(\mathbf{A}^{-1}\mathbf{y}-\boldsymbol{\mu})). \quad (5.6\text{-}15)$$

Can this formidable expression be put in the form of Equation 5.6-9? First we note that

$$[\det(\mathbf{K_{XX}})]^{1/2}|\det(\mathbf{A})| = [\det(\mathbf{A}\mathbf{K_{XX}}\mathbf{A}^T)]^{1/2}. \quad (5.6\text{-}16)$$

Next, factoring $\mathbf{A}$ inverse out of the first and last factors, and combining these terms with the inverse covariance matrix, we obtain

$$(\mathbf{A}^{-1}\mathbf{y}-\boldsymbol{\mu})^T\mathbf{K}_{\mathbf{XX}}^{-1}(\mathbf{A}^{-1}\mathbf{y}-\boldsymbol{\mu}) = (\mathbf{y}-\mathbf{A}\boldsymbol{\mu})^T(\mathbf{A}\mathbf{K_{XX}}\mathbf{A}^T)^{-1}(\mathbf{y}-\mathbf{A}\boldsymbol{\mu}). \quad (5.6\text{-}17)$$

But $\mathbf{A}\boldsymbol{\mu} \triangleq \boldsymbol{\beta} = E[\mathbf{Y}]$ and $\mathbf{A}\mathbf{K_{XX}}\mathbf{A}^T = E[(\mathbf{Y}-\boldsymbol{\beta})(\mathbf{Y}-\boldsymbol{\beta})^T] = \mathbf{K_{YY}}$. Hence Equation 5.6-15 can be rewritten as

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}[\det(\mathbf{K_{YY}})]^{1/2}}\exp[-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\beta})^T\mathbf{K}_{\mathbf{YY}}^{-1}(\mathbf{y}-\boldsymbol{\beta})]. \quad \blacksquare \quad (5.6\text{-}18)$$

The next question that arises quite naturally as an extension of the previous result is: Does $\mathbf{Y}$ remain a Normal random vector under more general (nontrivial) linear transformation? The answer is given by the following theorem, which is a generalization of Theorem 5.6-1.

Theorem 5.6-2 Let $\mathbf{X}$ be an n-dimensional Normal random vector with positive definite covariance matrix $\mathbf{K_{XX}}$ and mean vector $\boldsymbol{\mu}$. Let $\mathbf{A}_{mn}$ be an $m \times n$ matrix of rank m. Then the random vector generated by

$$\mathbf{Y} = \mathbf{A}_{mn}\mathbf{X}$$

has an m-dimensional Normal pdf with p.d. covariance matrix $\mathbf{K_{YY}}$ and mean vector $\boldsymbol{\beta}$ given, respectively, by

$$\mathbf{K_{YY}} \triangleq \mathbf{A}_{mn}\mathbf{K_{XX}}\mathbf{A}_{mn}^T \quad (5.6\text{-}19)$$

and

$$\boldsymbol{\beta} = \mathbf{A}_{mn}\boldsymbol{\mu}. \quad \blacksquare \quad (5.6\text{-}20)$$

The proof of this theorem is quite similar to the proof of Theorem 5.6-1; it is given by Miller in [5-6, p. 22].

Some examples involving transformations of Normal random variables are given below.

Example 5.6-1

(*transforming to independence*) A zero-mean Normal random vector $\mathbf{X} = (X_1, X_2)^T$ has covariance matrix $\mathbf{K_{XX}}$ given by

$$\mathbf{K_{XX}} = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}.$$

Find a transformation $\mathbf{Y} = \mathbf{CX}$ such that $\mathbf{Y} = (Y_1, Y_2)^T$ is a Normal random vector with uncorrelated (and therefore independent) components of unity variance.

Solution Write

$$E[\mathbf{YY}^T] = E[\mathbf{CXX}^T\mathbf{C}^T] = \mathbf{CK_{XX}C}^T = \mathbf{I}.$$

The last equality on the right follows from the requirement that the covariance of $\mathbf{Y}$, $\mathbf{K_{YY}}$, satisfies

$$\mathbf{K_{YY}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \triangleq \mathbf{I}.$$

From the previous discussion on whitening, the matrix $\mathbf{C}$ must be $\mathbf{C} = \Lambda^{-1/2}\mathbf{U}^T$, where $\Lambda^{-1/2}$ is the normalizing matrix

$$\Lambda^{-1/2} \triangleq \begin{bmatrix} \lambda_1^{-1/2} & 0 \\ 0 & \lambda_2^{-1/2} \end{bmatrix} \quad (\lambda_i, i = 1, 2 \text{ are eigenvalues of } \mathbf{K_{XX}})$$

and $\mathbf{U}$ is the matrix whose columns are the unit eigenvectors of $\mathbf{K_{XX}}$ (recall $\mathbf{U}^{-1} = \mathbf{U}^T$). From $\det(\mathbf{K_{XX}} - \lambda\mathbf{I}) = 0$, we find $\lambda_1 = 4$, $\lambda_2 = 2$. Hence

$$\mathbf{\Lambda} = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}, \qquad \mathbf{Z} = \mathbf{\Lambda}^{-1/2} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Next from

$$(\mathbf{K_{XX}} - \lambda_1\mathbf{I})\phi_1 = 0, \quad \text{with} \quad ||\phi_1|| = 1,$$

and

$$(\mathbf{K_{XX}} - \lambda_2\mathbf{I})\phi_2 = 0, \quad \text{with} \quad ||\phi_2|| = 1,$$

we find $\phi_1 = (1/\sqrt{2}, -1\sqrt{2})^T$, $\phi_2 = (1/\sqrt{2}, 1\sqrt{2})^T$. Thus,

$$\mathbf{U} = (\phi_1, \phi_2) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

and

$$\mathbf{C} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

As a check to see if $\mathbf{CK_{XX}C}^T$ is indeed an identity covariance matrix, we compute

$$\frac{1}{2} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In some situations we might want to generate correlated samples of a random vector $\mathbf{X}$ whose covariance matrix $\mathbf{K_{XX}}$ is not diagonal. From Example 5.6-1 we see that the transformation

$$\mathbf{X} = \mathbf{C}^{-1}\mathbf{Y}, \tag{5.6-21}$$

where $\mathbf{C} = \mathbf{Z}\mathbf{U}^T$ produces a Normal random vector whose covariance is $\mathbf{K_{XX}}$. Thus, one way of obtaining correlated from uncorrelated samples is to use the transformation given in Equation 5.6-21 on jointly independent computer-generated samples. This procedure is the reverse of what we did in Example 5.6-1.

Example 5.6-2

(*correlated Normal RVs*) Jointly Normal RVs X_1 and X_2 have joint pdf given by (See Equation 4.3-27 and the surrounding discussion in Section 4.3.)

$$f_{X_1X_2}(x_1, x_2) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2\sigma^2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\right).$$

Let the correlation coefficient ρ be -0.5. From X_1, X_2 find two jointly Normal RVs Y_1 and Y_2 such that Y_1 and Y_2 are independent. Avoid the trivial case of $Y_1 = Y_2 = 0$.

Solution Define $\mathbf{x} \triangleq (x_1, x_2)^T$ and $\mathbf{y} = (y_1, y_2)^T$. Then with $\rho = -0.5$, the quadratic in the exponent can be written as

$$x_1^2 + x_1x_2 + x_2^2 = \mathbf{x}^T \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mathbf{x} = ax_1^2 + (b+c)x_1x_2 + dx_2^2,$$

where the a, b, c, d are to be determined. We immediately find that $a = d = 1$ and—because of the real symmetric requirement—we find $b = c = 0.5$. We can rewrite $f_{X_1X_2}(x_1, x_2)$ in standard form as

$$f_{X_1X_2}(x_1, x_2) = \frac{1}{2\pi[\det(\mathbf{K_{XX}})]^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^T\, \mathbf{K}_{\mathbf{XX}}^{-1}\mathbf{x})\right),$$

whence

$$\mathbf{K}_{\mathbf{XX}}^{-1} = \frac{1}{\sigma^2(1-\rho^2)} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{4}{3\sigma^2} \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Our task is now to find a transformation that diagonalizes $\mathbf{K}_{\mathbf{XX}}^{-1}$. This will enable the joint pdf of Y_1 to Y_2 to be factored, thereby establishing that Y_1 and Y_2 are independent.

The factor $4/3\sigma^2$ affects the eigenvalues of $\mathbf{K}_{\mathbf{XX}}^{-1}$ but not the eigenvectors. To compute a set of orthonormal eigenvectors of $\mathbf{K}_{\mathbf{XX}}^{-1}$, we need only consider $\tilde{\mathbf{K}}_{\mathbf{XX}}^{-1}$ given by

$$\tilde{\mathbf{K}}_{\mathbf{XX}}^{-1} \triangleq \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

for which we obtain $\tilde{\lambda}_1 = 3/2$, $\tilde{\lambda}_2 = 1/2$. The corresponding unit eigenvectors are $\phi_1 = (1/\sqrt{2})(1, 1)^T$ and $\phi_2 = (1/\sqrt{2})(1, -1)^T$. Thus with

$$\tilde{\mathbf{U}} \triangleq \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

(normalization by $1/\sqrt{2}$ is not needed to obtain a diagonal covariance matrix so we dispense with these factors) we find that

$$\tilde{\mathbf{U}}^T\tilde{\mathbf{K}}_{\mathbf{XX}}^{-1}\tilde{\mathbf{U}} = \text{diag}(3, 1).$$

Hence a transformation that will work is[†]

$$\mathbf{Y} = \tilde{\mathbf{U}}^T\mathbf{X},$$

that is,

$$Y_1 = X_1 + X_2$$
$$Y_2 = X_1 - X_2.$$

To find $f_{Y_1Y_2}(y_1, y_2)$ we use Equation 3.4-21 of Chapter 3:

$$f_{Y_1Y_2}(y_1, y_2) = \sum_{i=1}^{n} f_{X_1X_2}(\mathbf{x}_i)/|J_i|,$$

where the $\mathbf{x}_i \triangleq (x_1^{(i)}, x_2^{(i)})^T$, $i = 1, \ldots, n$, are the n solutions to $\mathbf{y} - \tilde{\mathbf{U}}^T\mathbf{x} = 0$ and J_i is the Jacobian. There is only one solution ($n = 1$) to $\mathbf{y} - \tilde{\mathbf{U}}^T\mathbf{x} = 0$, which is

$$x_1 = \frac{y_1 + y_2}{2}$$
$$x_2 = \frac{y_1 - y_2}{2}$$

and, dispensing with subscripts there being only one root,

$$J = \det\left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right) = \det\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = -2.$$

Hence

$$\begin{aligned} f_{Y_1Y_2}(y_1, y_2) &= \frac{1}{2} f_{X_1X_2}\left(\frac{y_1 + y_2}{2}, \frac{y_1 - y_2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y_1^2}{2\sigma^2}\right] \cdot \frac{1}{\sqrt{2\pi\sigma'^2}} \exp\left[-\frac{y_2^2}{2\sigma'^2}\right], \end{aligned}$$

where $\sigma' \triangleq \sqrt{3}\sigma$.

[†]There is no requirement to whiten the covariance matrix as in Example 5.6-1. Also, diagonalizing $\mathbf{K}_{\mathbf{XX}}^{-1}$ is equivalent to diagonalizing $\mathbf{K}_{\mathbf{XX}}$.

Examples 5.6-1 and 5.6-2 are special cases of the following theorem:

Theorem 5.6-3 Let $\mathbf{X}$ be a Normal, zero-mean (for convenience) random vector with positive definite covariance matrix $\mathbf{K_{XX}}$. Then there exists a nonsingular $n \times n$ matrix $\mathbf{C}$ such that under the transformation

$$\mathbf{Y} = \mathbf{C}^{-1}\mathbf{X},$$

the components $Y_1, \ldots, Y_n$ of $\mathbf{Y}$ are independent and of unit variance.

Proof Let $\mathbf{C}^{-1} = \Lambda^{-1/2}\mathbf{U}^T$;

$$\text{then } \mathbf{K_{XX}} = \mathbf{C}\mathbf{C}^T. \quad \blacksquare$$

Example 5.6-3

(*generalized Rayleigh law*) Let $\mathbf{X} = (X_1, X_2, X_3)^T$ be a Normal random vector with covariance matrix

$$\mathbf{K_{XX}} = \sigma^2\mathbf{I}.$$

Compute the pdf of $R_3 \triangleq ||\mathbf{X}|| = \sqrt{X_1^2 + X_2^2 + X_3^2}$.

Solution The probability of the event $\{R_3 \leq r\}$ is the CDF $F_{R_3}(r)$ of R_3. Thus,

$$F_{R_3}(r) = \frac{1}{|(2\pi)^{3/2}[\sigma^2]^{3/2}]} \iiint_{\mathscr{I}} \exp\left[-\frac{1}{2\sigma^2}(x_1^2 + x_2^2 + x_3^2)\right] dx_1\, dx_2\, dx_3,$$

where $\mathscr{I} \triangleq \{(x_1, x_2, x_3)\colon \sqrt{x_1^2 + x_2^2 + x_3^2} \leq r\}$. Now let

$$x_1 \triangleq \xi \cos\phi$$

$$x_2 \triangleq \xi \sin\phi \cos\theta$$

$$x_3 = \xi \sin\phi \sin\theta,$$

that is, a rectangular-to-spherical coordinate transformation. The Jacobian of this transformation is $\xi^2 \sin\phi$. Using this transformation in the expression for $F_{R_3}(r)$, we obtain for $r > 0$

$$F_{R_3}(r) = \frac{1}{(2\pi)^{3/2}(\sigma^2)^{3/2}} \int_0^r \int_{\theta=0}^{2\pi} \int_{\phi=0}^{\pi} \exp\left[-\frac{\xi^2}{2\sigma^2}\right] \xi^2 \sin\phi\, d\xi\, d\theta\, d\phi$$

$$= \frac{4\pi}{(2\pi)^{3/2}[\sigma^2]^{3/2}} \int_0^r \xi^2 \exp\left[-\frac{\xi^2}{2\sigma^2}\right] d\xi.$$

To obtain $f_{R_3}(r)$, we differentiate $F_{R_3}(r)$ with respect to r. This yields

$$f_{R_3}(r) = \frac{2r^2}{\Gamma(\frac{3}{2})[2\sigma^2]^{3/2}} \exp\left[-\frac{r^2}{2\sigma^2}\right] \cdot u(r), \tag{5.6-22}$$

where $u(r)$ is the unit step and $\Gamma(3/2) = \sqrt{\pi}/2$. Equation 5.6-22 is an extension of the ordinary two-dimensional Rayleigh introduced in Chapter 2. The general n-dimensional Rayleigh is the pdf associated with $R_n \triangleq ||\mathbf{X}|| = \sqrt{X_1^2 + \ldots + X_n^2}$ and is given by

$$f_{R_n}(r) = \frac{2r^{n-1}}{\Gamma\left(\frac{n}{2}\right)[2\sigma^2]^{n/2}} \exp\left[-\frac{r^2}{2\sigma^2}\right] \cdot u(r). \tag{5.6-23}$$

The proof of Equation 5.6-23 requires the use of n-dimensional spherical coordinates. Such generalized spherical coordinates are well known in the mathematical literature [5-5, p. 9]. The demonstration of Equation 5.6-23 is left as a challenging problem.

5.7 CHARACTERISTIC FUNCTIONS OF RANDOM VECTORS

In Equation 4.7-1 we defined the CF of a random variable as

$$\Phi_X(\omega) \triangleq E[e^{j\omega X}].$$

The extension to random vectors is straightforward. Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be a real n-component random vector. Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ be a real n-component parameter vector. The CF of $\mathbf{X}$ is defined as

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) \triangleq E[e^{j\boldsymbol{\omega}^T\mathbf{X}}]. \tag{5.7-1}$$

The similarity to the scalar case is obvious. In the case of continuous random vectors, the actual evaluation of Equation 5.7-1 is done through

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x}) e^{j\boldsymbol{\omega}^T\mathbf{x}} d\mathbf{x}. \tag{5.7-2}$$

In Equation 5.7-2 we use the usual compact notation that $d\mathbf{x} = dx_1 \ldots dx_n$ and the integral sign refers to an n-fold integration. If $\mathbf{X}$ is a discrete random vector, $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ can be computed from the joint PMF as

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \sum_{-\infty}^{+\infty} P_{\mathbf{X}}(\mathbf{x}) e^{j\boldsymbol{\omega}^T\mathbf{x}}, \tag{5.7-3}$$

where the summation sign refers to an n-fold summation.
In both cases, we see that $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ is, except for a sign reversal in the exponent, the n-dimensional Fourier transform of $f_{\mathbf{X}}(\mathbf{x})$ or $P_{\mathbf{X}}(\mathbf{x})$. This being the case, we can recover for example the pdf by the inverse n-dimensional Fourier transform (again with a sign reversal). Thus,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \Phi_{\mathbf{X}}(\boldsymbol{\omega}) e^{-j\boldsymbol{\omega}^T\mathbf{x}} d\boldsymbol{\omega}. \tag{5.7-4}$$

The CF is very useful for computing joint moments. We illustrate with an example.

Example 5.7-1

(*finding mixed moment*) Let $\mathbf{X} \triangleq (X_1, X_2, X_3)^T$ and $\boldsymbol{\omega} \triangleq (\omega_1, \omega_2, \omega_3)^T$. Compute $E[X_1X_2X_3]$.

Solution Since

$$\Phi_{\mathbf{X}}(\omega_1, \omega_2, \omega_3) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathbf{x}}(x_1, x_2, x_3)e^{j[\omega_1 x_1+\omega_2 x_2+\omega_3 x_3]}\,dx_1\,dx_2\,dx_3,$$

we obtain by partial differentiation

$$\begin{aligned}&\frac{1}{j^3}\left.\frac{\partial^3\Phi_{\mathbf{X}}(\omega_1,\omega_2,\omega_3)}{\partial\omega_1\partial\omega_2\partial\omega_3}\right|_{\omega_1=\omega_2=\omega_3=0}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2x_3 f_{\mathbf{X}}(x_1,x_2,x_3)\,dx_1\,dx_2\,dx_3\\ &\triangleq E[X_1X_2X_3].\end{aligned}$$

Any moment—provided that it exists—can be computed by the method used in Example 5.7-1, that is, by partial differentiation. Thus,

$$E[X_1^{k_1}\dots X_n^{k_n}] = j^{-(k_1+\dots+k_n)}\left.\frac{\partial^{k_1+\dots+k_n}\Phi_{\mathbf{X}}(\omega_1,\dots,\omega_n)}{\partial\omega_1^{k_1}\dots\partial\omega_n^{k_n}}\right|_{\omega_1=\dots=\omega_n=0} \quad (5.7\text{-}5)$$

By writing

$$E[\exp(j\boldsymbol{\omega}^T\mathbf{X})] = E\left[\exp\left(j\sum_{i=1}^{n}\omega_i X_i\right)\right] = E\left[\prod_{i=1}^{n}\exp(j\omega_i X_i)\right]$$

and expanding each term in the product into a power series, we readily obtain the rather cumbersome formula

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \sum_{k_1=0}^{\infty}\dots\sum_{k_n=0}^{\infty} E[X_1^{k_1}\dots X_n^{k_n}]\frac{(j\omega_1)^{k_1}}{k_1!}\dots\frac{(j\omega_n)^{k_n}}{k_n!}, \quad (5.7\text{-}6)$$

which has the advantage of explicitly revealing the relationship between the joint CF and the joint moments of the $X_i, i = 1, \dots, n$. Of course Equation 5.7-6 has meaning only if

$$E[X_1^{k_1}\dots X_n^{k_n}]$$

exists for all values of the nonnegative integers $k_1, \dots, k_n$, and when the power series converge.

From Equation 5.7-2 observe the important CF properties:

Properties of CF of Random Vectors

1. $|\Phi_{\mathbf{X}}(\boldsymbol{\omega})| \leq \Phi_{\mathbf{X}}(\mathbf{0}) = 1$ and
2. $\Phi^*_{\mathbf{X}}(\boldsymbol{\omega}) = \Phi_{\mathbf{X}}(-\boldsymbol{\omega})$ ($*$ indicates conjugation).
3. All CFs of *subsets* of the components of $\mathbf{X}$ can be obtained once $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ is known.

The last property is readily demonstrated with the following example. Suppose $\mathbf{X} = (X_1, X_2, X_3)^T$ has CF† $\Phi_{\mathbf{X}}(\omega_1, \omega_2, \omega_3) = E[\exp j(\omega_1 X_1 + \omega_2 X_2 + \omega_3 X_3)]$. Then

$$\Phi_{X_1X_2}(\omega_1, \omega_2) = \Phi_{X_1X_2X_3}(\omega_1, \omega_2, 0)$$
$$\Phi_{X_1X_3}(\omega_1, \omega_3) = \Phi_{X_1X_2X_3}(\omega_1, 0, \omega_3)$$
$$\Phi_{X_1}(\omega_1) = \Phi_{X_1X_2X_3}(\omega_1, 0, 0).$$

As pointed out in Chapter 4, CFs are also useful in solving problems involving sums of independent RVs. Thus, suppose $\mathbf{X} = (X_1, \ldots, X_n)^T$, where the X_i are *independent* RVs with marginal pdf's $f_{X_i}(x_i), i = 1, \ldots, n$. The pdf of the sum

$$Z = X_1 + \ldots + X_n$$

can be obtained from

$$f_Z(z) = f_{X_1}(z) * \ldots * f_{X_n}(z). \tag{5.7-7}$$

However, the actual carrying out of the n-fold convolution in Equation 5.7-7 can be quite tedious. The computation of $f_Z(z)$ can be done more advantageously using CFs as follows. We have

$$\begin{aligned}\Phi_{\mathbf{Z}}(\boldsymbol{\omega}) &= E[e^{j\omega(X_1+\ldots+X_n)}] \\ &= \prod_{i=1}^{n} E\left[e^{j\omega X_i}\right] \\ &= \prod_{i=1}^{n} \Phi_{X_i}(\omega).\end{aligned} \tag{5.7-8}$$

In this development, line 2 follows from the fact that if $X_1, \ldots, X_n$ are n independent RVs, then $Y_i = g_i(X_i)$, $i = 1, \ldots, n$, will also be n independent RVs and $E[Y_1 \ldots Y_n] = E[Y_1] \ldots E[Y_n]$. The inverse Fourier transform of Equation 5.7-8 yields the pdf $f_Z(z)$. This approach works equally well when the X_i are discrete. Then the PMF and the discrete Fourier transform can be used. We illustrate this approach to computing the pdf's of sums of RVs with an example.

Example 5.7-2

(*i.i.d. Poison CF*) Let $\mathbf{X} = (X_1, \ldots, X_n)^T$, where the X_i, $i = 1, \ldots, n$ are i.i.d. Poisson RVs with Poisson parameter λ. Let $Z = X_1 + \ldots + X_n$. Then the individual PMFs are

†We use $\Phi_{\mathbf{X}}(\cdot)$ and $\Phi_{X_1X_2X_3}(\cdot)$ interchangeably if $\mathbf{X} = (X_1, X_2, X_3)^T$.

$$P_{X_i}(k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{5.7-9}$$

and

$$\begin{aligned}\Phi_{X_i}(\omega) &= \sum_{k=0}^{\infty} \frac{\lambda^k \exp(j\omega k)}{k!} e^{-\lambda} \\ &= e^{\lambda(\exp(j\omega)-1)}.\end{aligned} \tag{5.7-10}$$

Hence, by independence we obtain

$$\begin{aligned}\Phi_Z(\omega) &= \prod_{i=1}^{n} e^{\lambda(\exp(j\omega)-1)} \\ &= e^{n\lambda(\exp(j\omega)-1)}.\end{aligned} \tag{5.7-11}$$

Comparing Equation 5.7-11 with Equation 5.7-10 we see by inspection that $\Phi_Z(z)$ is the CF of the PMF

$$P_Z(k) = \frac{\alpha^k e^{-\alpha}}{k!}, \qquad k = 0, 1 \ldots, \tag{5.7-12}$$

where $\alpha \triangleq n\lambda$. Thus, the sum of n i.i.d. Poisson RVs is Poisson with parameter $n\lambda$.

The Characteristic Function of the Gaussian (Normal) Law

Let $\mathbf{X}$ be a real Gaussian (Normal) random vector with nonsingular covariance matrix $\mathbf{K_{XX}}$. Then from Theorem 5.6-3 both $\mathbf{K_{XX}}$ and $\mathbf{K_{XX}^{-1}}$ can be factored as

$$\mathbf{K_{XX}} = \mathbf{CC}^T \tag{5.7-13}$$

$$\mathbf{K_{XX}^{-1}} = \mathbf{DD}^T, \qquad \mathbf{D} \triangleq [\mathbf{C}^T]^{-1}, \tag{5.7-14}$$

where $\mathbf{C}$ and $\mathbf{D}$ are nonsingular. This observation will be put to good use shortly. The CF of $\mathbf{X}$ is by definition

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{n/2}[\det(\mathbf{K_{XX}})]^{1/2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{K_{XX}^{-1}}(\mathbf{x}-\boldsymbol{\mu})\right) \cdot \exp(j\boldsymbol{\omega}^T\mathbf{x})\, d\mathbf{x}. \tag{5.7-15}$$

Now introduce the transformation

$$\mathbf{z} \triangleq \mathbf{D}^T(\mathbf{x}-\boldsymbol{\mu}) \tag{5.7-16}$$

so that

$$\begin{aligned}\mathbf{z}^T\mathbf{z} &= (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{DD}^T(\mathbf{x}-\boldsymbol{\mu}) \\ &= (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{K_{XX}^{-1}}(\mathbf{x}-\boldsymbol{\mu}).\end{aligned} \tag{5.7-17}$$

The Jacobian of this transformation is $\det(\mathbf{D}^T) = \det(\mathbf{D})$. Thus under the transformation in Equation 5.7-16, Equation 5.7-15 becomes

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \frac{\exp(j\boldsymbol{\omega}^T\boldsymbol{\mu})}{(2\pi)^{n/2}[\det(\mathbf{K}_{\mathbf{XX}})]^{1/2}|\det(\mathbf{D})|}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right)\cdot\exp(j\boldsymbol{\omega}^T(\mathbf{D}^T)^{-1}\mathbf{z})\,d\mathbf{z}. \tag{5.7-18}$$

We can complete the squares in the integrand as follows:

$$\exp[-\{\tfrac{1}{2}[\mathbf{z}^T\mathbf{z} - 2j\boldsymbol{\omega}^T(\mathbf{D}^T)^{-1}\mathbf{z}]\}] = \exp(-\tfrac{1}{2}\boldsymbol{\omega}^T(\mathbf{D}^T)^{-1}(\mathbf{D})^{-1}\boldsymbol{\omega})$$
$$\cdot\exp(-\tfrac{1}{2}||\mathbf{z} - j\mathbf{D}^{-1}\boldsymbol{\omega}||^2). \tag{5.7-19}$$

Equations 5.7-18 and 5.7-19 will be greatly simplified if we use the following results: (a) If $\mathbf{K}_{\mathbf{XX}}^{-1} = \mathbf{D}\mathbf{D}^T$, then $\mathbf{K}_{\mathbf{XX}} = [\mathbf{D}^T]^{-1}\mathbf{D}^{-1}$; (b) $\det(\mathbf{K}_{\mathbf{XX}}^{-1}) = \det(\mathbf{D})\det(\mathbf{D}^T) = [\det(\mathbf{D})]^2 = [\det(\mathbf{K}_{\mathbf{XX}})]^{-1}$. Hence $|\det(\mathbf{D})|^{-1} = [\det(\mathbf{K}_{\mathbf{XX}})]^{1/2}$. It then follows that

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \exp\left(j\boldsymbol{\omega}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\omega}^T\mathbf{K}_{\mathbf{XX}}\boldsymbol{\omega}\right)\cdot\frac{1}{(2\pi)^{n/2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}||\mathbf{z}-j\mathbf{D}^{-1}\boldsymbol{\omega}||^2}\,d\mathbf{z}.$$

Finally we recognize that the n-fold integral on the right-hand side is the product of n identical integrals of one-dimensional Gaussian densities, each of unit variance. Hence the value of the integral is merely $(2\pi)^{n/2}$, which cancels the factor $(2\pi)^{-n/2}$ and yields the CF for the Normal random vector:

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \exp[j\boldsymbol{\omega}^T\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{\omega}^T\mathbf{K}_{\mathbf{XX}}\boldsymbol{\omega}], \tag{5.7-20}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$, and $\mathbf{K}_{\mathbf{XX}}$ is the covariance. We observe in passing that $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ has a multidimensional complex Gaussian form as a function of $\boldsymbol{\omega}$. Thus, the Gaussian pdf has mapped into a Gaussian CF, a result that should not be too surprising since we already know that the one-dimensional Fourier transform maps a Gaussian function into a Gaussian function.

Similarly the joint MGF for a random vector $\mathbf{X} = (X_1, \ldots, X_N)^T$ is defined as

$$M_{\mathbf{X}}(\boldsymbol{t}) \triangleq E\left[\exp\sum_{i=1}^{N} t_i X_i\right]$$
$$= \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\cdots\sum_{k_N=0}^{\infty}\frac{t_1^{k_1}}{k_1!}\cdots\frac{t_N^{k_N}}{k_N!}E\left[X_1^{k_1}X_2^{k_2}\ldots X_N^{k_N}\right]$$

from which joint moments can be computed analogously to the CF case.

SUMMARY

In this chapter we studied the calculus of multiple RVs. We found it convenient to organize multiple RVs into random vectors and treat these as single entities. We found that when i.i.d.

random variables are *ordered*, many probabilistic results can be derived without specifying the underlying distributions. In Section 5.3, we derived, among others, the distribution of probability area (the area under the pdf between order samples) and the moments of such probability areas. We shall see in subsequent chapters that ordered random variables play important roles in a branch of statistics called *distribution-free* or *robust statistics*. Because in practice it is often difficult to describe the joint probability law of n RVs, we argued that in the case of random vectors we often settle for a less complete but more available characterization than that furnished by the pdf (PMF). We focused on the characterizations furnished by the lower order moments, especially the mean and covariance. In particular, because of the great importance of covariance matrices in signal processing, communication theory, pattern recognition, multiple regression analysis, and other areas of engineering and science, we made use of numerous results from matrix theory and linear algebra to reveal the properties of these matrices.

We discussed the multidimensional Gaussian (Normal) law and CFs of random vectors. We demonstrated that under linear transformations Gaussian random vectors map into Gaussian random vectors. We showed how to derive a transformation that can convert correlated RVs into uncorrelated ones. The CF of random vectors in general was defined and shown to be useful in computing moments and solving problems involving the sums of independent RVs; these assertions were illustrated with examples. Finally, using vector and matrix techniques we derived the CF for the Gaussian random vector and showed that it too had a Gaussian shape.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

5.1 Let X_1, X_2 and X_3 be independent standard normal random variables. Let

$$Y_1 = X_1 + X_2 + X_3$$

$$Y_2 = X_1 - X_2$$

$$Y_3 = X_2 - X_3$$

Determine the joint pdf of Y_1, Y_2, Y_3.

5.2 Let $B_i, i = 1, \ldots, n$, be n disjoint and exhaustive events. Show that the CDF of $\boldsymbol{X}$ can be written as

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{n} F_{\mathbf{X}|B_i}(\mathbf{x}|B_i)P[B_i].$$

5.3 Two Gaussian random variables X_1 and X_2 have zero means and variances $\sigma_{x1}^2 = 4$ and $\sigma_{x2}^2 = 9$. Their covariance is $K_{X1X2} = 3$. If X_1 and X_2 are linearly transformed to new variables Y_1 and Y_2 according to $Y_1 = X_1 - 2X_2$ and $Y_2 = 3X_1 + 4X_2$, find the means, variances and covariance of Y_1 and Y_2.

5.4 Let X_1, X_2, X_3 be three standard Normal RV's. For $i = 1, 2, 3$ let $Y_i \in \{X_1, X_2, X_3\}$ such that $Y_1 < Y_2 < Y_3$ i.e. the ordered—by—signed magnitude of the X_i. Compute the joint pdf $f_{Y_1Y_2Y_3}(y_1, y_2, y_3)$.

5.5 In Problem 5.4 compute the CDF $F_{R_1}(y)$, for $i = 1, 2, 3$ and plot the result.

5.6 In Section 5.4 we introduced the RVs Z_1 and Z_n. Show that the joint pdf of Z_1 and Z_n is given by Equation 5.3-7.

5.7 Consider the RVs $V_{\text{ln}} \triangleq Z_n - Z_1$, $W = Z_n$. Show that the joint pdf $f_{V_{\text{ln}}}W(v, w) = n(n-1)v^{n-2}$, for $0 < w - v < w < 1$ and zero else.

5.8 From the results of the previous problem, show that $f_{v_{\text{ln}}}(v) = n(n-1)v^{n-2}(1-v)$, for $0 < v < 1$, $n \geq 2$ and zero else.

5.9 Show that the area under $f_{Z_1Z_2Z_3}(z_1, z_2, z_3) = 3!$ with $0 < z_1 < z_2 < z_3 < 1$ is unity.

5.10 Compute the beta CDF for $n = 2$, $\beta = 0$; $n = 2$, $\beta = 0$.

5.11 Derive Equations 5.3-11, 5.3-12, 5.3-13.

5.12 Use Excel or a similar computer program to generate curves of the beta CDF for $n = 15, 20, 30$. Describe what seems to happening as $n \to \infty$.

5.13 Derive Equation 5.3-14.

5.14 Show that, on the average, n ordered random variables divide that total area under $f_X(x)$ into $n + 1$ equal parts.

5.15 Show that any matrix $\mathbf{M}$ generated by an outer product of two vectors, that is, $\mathbf{M} = \mathbf{X}\mathbf{X}^T$, has rank at most unity. Explain why $\mathbf{R} \triangleq E[\mathbf{X}\mathbf{X}^T]$ can be of full rank.

5.16 Let $\{X_i, i = 1, \ldots, n\}$ be n i.i.d. observation on X and let $\{Y_i, i = 1, \ldots, n\}$ be the associated order statistics. Show that $F_{Y_n}(y) = F_X^n(y)$.

5.17 Let $\{X_i, i = 1, \ldots, n\}$ be n i.i.d. observation on X and let $\{Y_i, i = 1, \ldots, n\}$ be the associated order statistics. Show that $F_{Y_1} = 1 - (1 - F_X(y))^n$.

5.18 Let $X = \cos\Theta$ and $Y = \sin\Theta$ where Θ is uniformly distributed in $(0, 2\pi)$. Determine whether X and Y are independent or not. Verify whether X and Y are uncorrelated.

5.19 Show that the two RVs X_1 and X_2 with joint pdf

$$f_{X_1X_2}(x_1, x_2) = \begin{cases} \frac{1}{16}, & |x_1| < 4, \quad 2 < x_2 < 4 \\ 0, & \text{otherwise} \end{cases}$$

are independent and orthogonal.

5.20 Let $\bar{X} = (X_1 X_2)$ consist of two unit-variance uncorrelated random variables. Find the matrix A such that $Y = AX$ has the covariance matrix

$$K = \sigma^2 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ where } \quad |\rho| < 1$$

5.21 Two random variables X and Y have the joint characteristic function $\phi_{X,Y}(w_1, w_2) = exp[-2w_1^2 - 8w_2^2]$
Show that X and Y are both zero-mean random variables and that they are uncorrelated.

5.22 Let $\mathbf{X}_i, i = 1, \ldots, n$, be n mutually uncorrelated random vectors with $E[\mathbf{X}_i] = \boldsymbol{\mu}_i$, $i = 1, \ldots, n$. Show that

$$E\left[\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_i)\sum_{j=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu}_j)^T\right] = \sum_{i=1}^{n}\mathbf{K}_i,$$

where $\mathbf{K}_i \triangleq E[(\mathbf{X}_i - \boldsymbol{\mu}_i)(\mathbf{X}_i - \boldsymbol{\mu}_i)^T]$.

5.23 The vector of random variables (X, Y, Z) is jointly Gaussian with zero means and the covariance matrix

$$K = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix}$$

Find the bivariate density of (X, Y).

5.24 (a) Let a vector $\mathbf{X}$ have $E[\mathbf{X}] = \mathbf{0}$ with covariance $\mathbf{K_{XX}}$ given by

$$\mathbf{K_{XX}} = \begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 4 \end{bmatrix}.$$

Find a linear transformation $\mathbf{C}$ such that $\mathbf{Y} = \mathbf{CX}$ will have

$$\mathbf{K_{YY}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Is $\mathbf{C}$ a unitary transformation?

(b) Consider the two real symmetric matrices $\mathbf{A}$ and $\mathbf{A}'$ given by

$$\mathbf{A} \triangleq \begin{bmatrix} a & b \\ b & c \end{bmatrix} \qquad \mathbf{A}' \triangleq \begin{bmatrix} a' & b' \\ b' & c' \end{bmatrix}.$$

Show that when $a = c$ and $a' = c'$, the product $\mathbf{AA}'$ is real symmetric. More generally, show that if $\mathbf{A}$ and $\mathbf{A}'$ are any real symmetric matrices, then $\mathbf{AA}'$ will be symmetric if $\mathbf{AA}' = \mathbf{A}'\mathbf{A}$.

(K. Fukunaga [5-8, p. 33].) Let $\mathbf{K}_1$ and $\mathbf{K}_2$ be positive definite covariance matrices and form

$$\mathbf{K} = a_1\mathbf{K}_1 + a_2\mathbf{K}_2, \qquad \text{where } a_1, a_2 > 0.$$

5.25 Let $\mathbf{A}$ be a transformation that achieves

$$\mathbf{A}^T\mathbf{K}\mathbf{A} = \mathbf{I} \qquad \mathbf{A}^T\mathbf{K}_1\mathbf{A} = \boldsymbol{\Lambda}^{(1)} = \text{diag}(\lambda_1^{(1)}, \ldots, \lambda_n^{(1)}).$$

(a) Show that $\mathbf{A}$ satisfies

$$\mathbf{K}^{-1}\mathbf{K}_1\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}^{(1)}.$$

(b) Show that $\mathbf{A}^T\mathbf{K}_2\mathbf{A} \triangleq \mathbf{\Lambda}^{(2)}$ is also diagonal, that is, $\mathbf{\Lambda}^{(2)} \triangleq \text{diag}(\lambda_1^{(2)}, \ldots, \lambda_n^{(2)})$.

(c) Show that $\mathbf{A}^T\mathbf{K}_1\mathbf{A}$ and $\mathbf{A}^T\mathbf{K}_2\mathbf{A}$ share the same eigenvectors.

(d) Show that the eigenvalues of $\mathbf{\Lambda}^{(2)}$ are related to the eigenvalues of $\mathbf{\Lambda}^{(1)}$ as

$$\lambda_i^{(2)} = \frac{1}{a_2}[1 - a_1\lambda_i^{(1)}]$$

and therefore are in inverse order from those of $\mathbf{\Lambda}^{(1)}$.

5.26 (J. A. McLaughlin [5-9].) Consider the m vectors $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})^T$, $i = 1, \ldots, m$, where $n > m$. Consider the $n \times n$ matrix $\mathbf{S} = \frac{1}{m}\sum_{i=1}^m \mathbf{X}_i\mathbf{X}_i^T$.

(a) Show that with $\mathbf{W} \triangleq (\mathbf{X}_1 \ldots \mathbf{X}_m)$, $\mathbf{S}$ can be written as

$$\mathbf{S} = \frac{1}{m}\mathbf{W}\mathbf{W}^T.$$

(b) What is the maximum rank of $\mathbf{S}$?

(c) Let $\mathbf{S}' \triangleq \frac{1}{m}\mathbf{W}^T\mathbf{W}$. What is the size of $\mathbf{S}'$? Show that the first m nonzero eigenvalues of $\mathbf{S}$ can be computed from

$$\mathbf{S}'\mathbf{\Phi} = \mathbf{\Phi}\mathbf{\Lambda},$$

where $\mathbf{\Phi}$ is the eigenvector matrix of $\mathbf{S}'$ and $\mathbf{\Lambda}$ is the matrix of eigenvalues. What are the relations between the eigenvectors and eigenvalues of $\mathbf{S}$ and $\mathbf{S}'$?

(d) What is the advantage of computing the eigenvectors from $\mathbf{S}'$ rather than $\mathbf{S}$?

5.27 (a) Let $\mathbf{K}$ be an $n \times n$ covariance matrix and let $\Delta\mathbf{K}$ be a real symmetric perturbation matrix. Let λ_i, $i = 1, \ldots, n$, be the eigenvalues of $\mathbf{K}$ and $\boldsymbol{\phi}_i$ the associated eigenvectors. Show that the first-order approximation to the eigenvalues λ_i' of $\mathbf{K} + \Delta\mathbf{K}$ yields

$$\lambda_i' = \boldsymbol{\phi}_i^T(\mathbf{K} + \Delta\mathbf{K})\boldsymbol{\phi}_i, \qquad i = 1, \ldots, n.$$

(b) Show that the first-order approximation to the eigenvectors is given by

$$\Delta\boldsymbol{\phi}_i = \sum_{j=1}^n b_{ij}\boldsymbol{\phi}_j,$$

where $b_{ij} = \boldsymbol{\phi}_j^T\Delta\mathbf{K}\boldsymbol{\phi}_i/(\lambda_i - \lambda_j)$ $i \neq j$ and $b_{ii} = 0$.

5.28 Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of a real symmetric matrix $\mathbf{M}$. For $i \geq 2$, let $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{i-1}$ be mutually orthogonal unit eigenvectors belonging to $\lambda_1, \ldots, \lambda_{i-1}$. Prove that the *maximum value* of $\mathbf{u}^T\mathbf{M}\mathbf{u}$ subject to $||\mathbf{u}|| = 1$ and $\mathbf{u}^T\boldsymbol{\phi}_1 = \ldots = \mathbf{u}^T\boldsymbol{\phi}_{i-1} = 0$ is λ_i, that is, $\lambda_i = \max(\mathbf{u}^T\mathbf{M}\mathbf{u})$.

5.29 Let $\mathbf{X} = (X_1, X_2, X_3)^T$ be a random vector with $\boldsymbol{\mu} \triangleq E[\mathbf{X}]$ given by

$$\boldsymbol{\mu} = (5, -5, 6)^T$$

and covariance given by

$$\mathbf{K} = \begin{bmatrix} 5 & 2 & -1 \\ 2 & 5 & 1 \\ -1 & 0 & 4 \end{bmatrix}.$$

Calculate the mean and variance of

$$Y = \mathbf{A}^T\mathbf{X} + B,$$

where $\mathbf{A} = (2, -1, 2)^T$ and $B = 5$.

5.30 Two jointly Normal RVs X_1 and X_2 have joint pdf given by

$$f_{X_1X_2}(x_1, x_2) = \frac{2}{\pi\sqrt{7}} \exp[-\tfrac{8}{7}(x_1^2 + \tfrac{3}{2}x_1x_2 + x_2^2)].$$

Find a nontrivial transformation $\mathbf{A}$ in

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \mathbf{A} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

such that Y_1 and Y_2 are independent. Compute the joint pdf of Y_1, Y_2.

5.31 Show that if $\mathbf{X} = (X_1, \ldots, X_n)^T$ has mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and covariance

$$\mathbf{K} = \{K_{ij}\}_{n \times n},$$

then the scalar RV Y given by

$$Y \triangleq p_1X_1 + \ldots + p_nX_n$$

has mean

$$E[Y] = \sum_{i=1}^{n} p_i\mu_i$$

and variance

$$\sigma_Y^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j K_{ij}.$$

5.32 Compute the joint characteristic function of $\mathbf{X} = (X_1, \ldots, X_n)^T$, where the $X_i, i = 1, \ldots, n$, are mutually independent and identically distributed Cauchy RVs, that is,

$$f_{X_i}(x) = \frac{\alpha}{\pi(x^2 + \alpha^2)}.$$

Use this result to compute the pdf of $Y = \sum_{i=1}^{n} X_i$.

5.33 Suppose that U and V are independent, zero-mean, unit variance Gaussian random variables. Let $X = U + V,\ Y = 2U + V$
Find the joint characteristic function of X and Y and find $E(XY)$.

5.34 Let $\mathbf{X} = (X_1, \ldots, X_4)$ be a Gaussian random vector with $E[\mathbf{X}] = 0$. Show that

$$E[X_1X_2X_3X_4] = K_{12}K_{34} + K_{13}K_{24} + K_{14}K_{23},$$

where the K_{ij} are elements of the covariance matrix $\mathbf{K} = \{K_{ij}\}_{4\times 4}$ of $\mathbf{X}$.

5.35 Let the joint pdf of X_1, X_2, X_3 be given by $f_{\mathbf{X}}(x_1, x_2, x_3) = 2/3 \cdot (x_1 + x_2 + x_3)$ over the region $S = \{(x_1, x_2, x_3) : 0 < x_i \leq 1, i = 1, 2, 3\}$ and zero elsewhere. Compute the covariance matrix and show that the random variables X_1, X_2, X_3, although not independent, are essentially uncorrelated.

5.36 Let X_1, X_2 be jointly Normal, zero-mean random variables with covariance matrix

$$\mathbf{K} = \begin{bmatrix} 2 & -1.5 \\ -1.5 & 2 \end{bmatrix}.$$

Find a whitening transformation for $\mathbf{X} = (X_1X_2)^T$. Write a MATLAB program to show a scatter diagram, that is, x_2 versus x_1 where the latter are realizations of X_2, X_1, respectively. Do this for the whitened variables as well. Choose between a hundred and a thousand realizations.

5.37 (linear transformations) Let $Y_k = \sum_{j=1}^{n} a_{kj}X_j, k = 1, \ldots, n$, where the a_{kj} are real constants, the matrix $\mathbf{A} = [a_{ij}]_{N\times N}$ is nonsingular, and the $\{X_j\}$ are random variables. Let $\mathbf{B} = \mathbf{A}^{-1}$. Show that the pdf of $\mathbf{Y}$, $f_{\mathbf{Y}}(y_1, \ldots, y_n)$ is given by

$$f_{\mathbf{Y}}(y_1, \ldots, y_n) = |\det \mathbf{B}| f_{\mathbf{X}}(x_1^*, \ldots, x_n^*), \text{ where } x_i^* = \sum_{k=1}^{n} b_{ik}y_k \text{ for } i = 1, \ldots, n.$$

5.38 (auxiliary variables) Let $Y_1 = \sum_{i=1}^{n} X_i$ and $Y_2 = \sum_{i=2}^{n} X_i$. Compute the joint pdf, $f_{Y_1Y_2}(y_1, y_2)$, by introducing the auxiliary variables $Y_k = \sum_{i=k}^{n} X_i$, $k = 3, \ldots, n$, and integrating over the range of each auxiliary RV. Show the $f_{\mathbf{Y}}(y_1, \ldots, y_n) = f_{\mathbf{X}}(y_1 - y_2, \ldots, y_{n-1} - y_n, y_n)$. (This problem and the previous are adapted from Example 4.9, p. 190, in *Probability and Stochastic Processes for Engineers*, C. W. Helstrom, Macmillan, 1984).

REFERENCES

5-1. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edition. New York: John Wiley, 1971.
5-2. W. B. Davenport, Jr., *Probability and Random Processes*, New York: McGraw-Hill, p. 99.
5-3. A. Papoulis, and S. U. Pillai *Probability, Random Variables, and Stochastic Processes.* New York: McGraw-Hill, 4th Ed, 2002.
5-4. B. Saleh, *Photoelectron Statistics.* New York: Springer-Verlag, 1978, Chapter 5.
5-5. R. G. Gallagher, *Information Theory and Reliable Communications.* New York: John Wiley, 1968.
5-6. P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Part 1. New York: McGraw-Hill, 1953, p. 279.
5-7. G. A. Korn and T. S. Korn, *Mathematical Handbook for Scientists and Engineers.* New York: McGraw-Hill, 1961.
5-8. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd edition. New York: John Wiley, 1957.
5-9. K. Fukunaga, Introduction to Statisticak Pattern recognition. 2nd edition. New York: Academic 1960.
5-10. J.A. McLaughlin and J. Raviv, "Nth Order Autocorrelations in Pattern Recognition, Information and Control", 12, pp. 121-142, Chapter 2, 1968 .

ADDITIONAL READING

Cooper, G. R. and C. D. McGillem, *Probabilistic Methods of Signal and System Analysis*, 3rd edition. New York: Holt, Rinehart and Winston, 1999.

Peebles, P. Z. Jr., *Probability, Random Variables, and Random Signal Principles*, 4th edition. New York: McGraw-Hill, 2001.

Leon-Garcia, A., *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd edition. Reading, MA: Prentice Hall, 2008.

Helstrom, C. W., *Probability and Stochastic Processes for Engineers*, 2nd edition. New York: Macmillan, 1991.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd edition. New York: McGraw-Hill, 1991.

Scheaffer, R. L., *Introduction to Probability and Its Applications.* Belmont, CA: Duxbury, 1990.

Viniotis, Y., *Probability and Random Processes for Electrical Engineers.* New York: McGraw-Hill, 1998.

Yates, R. D. and D. J. Goodman, *Probability and Stochastic Processes*, 2nd edition, New York: Wiley, 2004.

6 Statistics: Part 1 Parameter Estimation

6.1 INTRODUCTION

Statistics, which could equally be called applied probability, is a discipline that applies the *principles of probability* to actual data. Two key areas of statistics are *parameter estimation* and *hypothesis testing.* In *parameter estimation*, we use real-world data to estimate parameters such as the mean, standard deviation, variance, covariance, probabilities, and distributions. In *hypothesis testing* we use real-world data to make rational decisions, if possible, in a probabilistic environment. We leave the topic of hypothesis testing for Chapter 7.

We recall that *probability* is a mathematical theory based on axioms and definitions and its main results are theorems, corollaries, relationships, and models. While *probability* enables us to model and solve a wide class of problems, the solutions to these problems often assume knowledge that is not readily available in the real world. For example, suppose we are given that $X{:}N(\mu, \sigma^2)$ and we wish to compute the probability of the event $E = \{-1 \leq X \leq +1\}$. We do this easily and obtain $F_{SN}((1-\mu)/\sigma) - F_{SN}((-1-\mu)/\sigma)$. However, in the real world how would we determine the parameters μ, σ? For that matter, how would we even determine that this is a Gaussian problem? In earlier chapters we used important parameters such as μ_X the average or expected value of a random variable RV X; σ_X, the standard deviation of X; σ_X^2, the variance of X; $E[XY]$, the correlation of two RVs X and Y; and others. We estimate these quantities in the real world using so called *estimators*, which are functions of RVs. What are the features of a good estimator? How do we choose among different estimators for the same parameter? What strong statements can we make regarding how "near" the estimate is to the true but unknown value?

Much has been written about parameter estimation but the subject is not exhausted, as witnessed by the large number of research articles in the archival literature devoted to the subject. There are several excellent books (e.g., [6-1, 6-2]) on statistics and parameter estimation with an "engineering" flavor, and a plethora of expository material on the internet.

Example 6.1-1

(*is a coin fair?*) Suppose you are involved in a game involving a coin and you would like to know if the coin is fair. You flip the coin and observe a *head* (H). What conclusion can you come to? Other than concluding that the coin is not restricted from coming up *heads* there isn't much else that you can conclude. Now you repeat the experiment and observe a *tail* (T) on the next toss. Can you conclude that the coin is fair? That would be a highly risky conclusion. Suppose that in 10 tosses you observe the sequence {H, T, T, H, H, T, H, T, H, T}. Based on the observations and using the frequency interpretation of probability, we might conclude that $P[\text{H}] = n_{\text{H}}/n = 5/10 = 0.5$ and thus that the coin is fair but you still cannot be certain. On the other hand, if you observe the sequence {H, T, H, H, H, H, T, H, H, H} you might be tempted to conclude that the coin is biased toward coming up *heads* but even here you can't be certain. Is there a quantitative way of describing our uncertainty (or certainty)? In what follows we introduce some ideas that will help to answer this question.

Independent, Identically Distributed (i.i.d.) Observations

In the coin tossing experiment described above, upon tossing a coin we can define a generic RV X as

$$X \triangleq \begin{cases} 1, & \text{if a head shows up,} \\ 0, & \text{if a tail shows up.} \end{cases}$$

If we toss the coin n times, we define a sequence of RVs $X_i, i = 1, \ldots, n$, which are called *independent, identically, distributed* (i.i.d.) *observations*. The collection of these i.i.d. observations $\{X_i; i = 1, \ldots, n\}$ is called a *random sample of size* n from X. In some situations, X is more aptly called *a population*; but the set of observations on X is still called a *random sample of size n*. The X_i in this example happen to be Bernoulli RVs but in general they could have any distributions as long as they all share the same CDF, pdf, or PMF and each observation is unaffected by the outcome of the previous distribution.

We have already introduced the idea of i.i.d. RVs in connection with our discussion of the Central Limit Theorem but elaborate on them some more here because of their extraordinary importance in statistics. The observations are *independent* because, in this case, subsequent tosses are not influenced in any way by the outcomes of previous tosses or future tosses. More precisely, in terms of the joint probability mass function(s) (PMF) of $X_i, i = 1, \ldots, n$:

$$P_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n) = P_{X_1}(x_1) P_{X_2}(x_2) \cdots P_{X_n}(x_n).$$

They are *identically distributed* because we are using the same coin in all the tosses and the coin is assumed unaffected by the experiment. More precisely:

$$P_{X_1}(x) = P_{X_2}(x) = \cdots = P_{X_n}(x) \triangleq P_X(x), \quad -\infty < x < \infty.$$

When we deal with continuous random variables the property i.i.d. implies:

$$f_{X_1X_2\cdots X_n}(x_1,x_2,\cdots,x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$$
$$f_{X_1}(x) = f_{X_2}(x) = \cdots = f_{X_n}(x) \triangleq f_X(x), -\infty < x < \infty.$$

The idea of i.i.d. observations is counterintuitive for many readers. For example, a coin—judged fair by all physical and previous statistical tests—is tossed and comes up heads nine times in a row; surely some readers will expect the coin to come up tails on the tenth toss to "balance things out." But the coin has no memory of its past history and on the tenth toss it is as likely to come up heads as tails.[†]

Example 6.1-2

(*failure of identically distributed condition*) We study the arrival rates of customers at a barbershop. To that end we partition the workday (7 am to 3 pm) into 16 half-hour intervals and count the number of arrivals in each interval. Let $X_i, i = 1,\ldots,16$, denote the number of arriving customers in the ith interval. Here the X_i are *not* i.i.d. (failure of the "identically distributed" requirement). We expect more arrivals in the early morning, before people must report to their jobs, than at other times in the day except possibly during the lunch break.

Example 6.1-3

(*biased random sampling*) A breakfast food company that produces BranPellets™ cereal intends to show that eating BranPellets™ will result in weight loss. To that end the company hires a pollster to poll those who have attempted to lose weight by eating BranPellets™. The pollster begins by randomly selecting from the pool of BranPellets™ eaters but when the results do not seem to confirm that eating BranPellets™ results in weight loss, the pollster confines the polling to the sub-group of people of average or less-than average weight. With X_i denoting the weight loss of the ith person polled after three months of eating BranPellets™, we note that the set of $\{X_i\}$ obtained by fair polling are unlikely to be distributed by the same law as the set of $\{X_i\}$ obtained by biased polling. Incidentally, we could formulate this as a hypothesis testing problem by formulating the hypothesis that eating BranPellets™ will result in weight loss versus the alternative that eating BranPellets™ *will not* result in weight loss.

Example 6.1-4

(*non-independent sequences*) A conservative gambler plays n rounds of blackjack. He starts with a stash of \$100 and bets only \$1 at each round. Let X_i denote the value of his stash at the ith play. Are the $X_i, i = 1,\ldots,n$, an independent sequence? Clearly $X_{i+1} = X_i \pm 1$ hence the X_i are not mutually independent; for example, $P[X_i = 10, X_{i+1} = 12] = 0$, although taken separately neither probability needs to be zero. Let Y_i denote the gambler's win (or loss) on the ith play. Then $Y_i = \pm 1$. Are the $Y_i, i = 1,\ldots,n$ an independent sequence? The answer is yes[†] because the outcome of the ith play has no memory of the past or future and therefore cannot be affected by it.

[†]However, if in a large number of tosses there are many more heads than tails, the assumption that the coin is fair needs to be re-examined. Here hypothesis testing (Chapter 7) is useful in making a stronger statement than the coin is "probably fair" or "probably unfair."

[†]Several assumptions are at play here, among them that the dealer plays fairly and that the gambler doesn't change strategy as a result of his wins or losses.

Example 6.1-5

(*review of joint versus sum probabilities*) We make three i.i.d. observations on a zero/one Bernoulli RV X and call these X_1, X_2, X_3[‡]. The PMF of X is $P_X(x) = p^x q^{1-x}, p+q=1$, $x = 0, 1$. The joint PMF of the observations is

$$P_{X_1X_2X_3}(x_1, x_2, x_3) = p^{x_1+x_2+x_3} q^{3-(x_1+x_2+x_3)}.$$

Note that this is different from the PMF of the sum $Y \triangleq X_1 + X_2 + X_3$, which is binomial with PMF $P_Y(k) = b(k; 3, p) = \binom{3}{k} p^k q^{3-k}$.

Estimation of Probabilities

Suppose that, based on observations, we *estimate* that the probability[§] of an event E is $\hat{P}[E] = n_E/n = 0.44$. Here n is the sample size and n_E is the number of times the event E is observed. How close is 0.44 to the "true" probability of the event? The "true" probability of an event is often beyond our means to acquire. Suppose a medical researcher wants to know the proportion (probability $\times$ 100) that his patient's red blood cells are undersized. The true proportion could, hypothetically, be obtained by counting all the undersized cells among all red blood cells in the patient's body and forming the ratio of the former to the latter. Of course this isn't done. Nevertheless an excellent estimate can be obtained by counting the cells in few drops of blood. As another example, suppose one of the states in the United States has a county with 343,065 registered voters and 144,087 have voted Republican. Then the true probability that a person in this county, picked at random, has voted Republican is 0.42. However, the cost of polling 343,065 voters may be prohibitive (or impossible in the time allowed) and pollster may have to make predictions with much smaller random samples. Thus, suppose that pollsters do a random sampling of 512 voters and find that 225 voters have voted Republican. Then the estimated probability of Republican voters is 0.44. Notice that if the sample size is small enough, the estimate of Republican voters can be almost any number between zero and one. For example if we poll only two voters and they both voted Republican, our estimate of the probability of Republican voters would be one! But this estimate would be completely unreliable! On the other hand, if we could say something like "with a near-certain probability of 0.98 the estimated probability of a Republican voter is between 0.42 and 0.46" then we have would have made a "hard" statement about the percentage of Republican voters. The probability 0.98 is a hard number because we can be nearly certain that the percentage of Republican voters is between 42 and 44 percent. Thus, the estimated probability of Republican voters is a "soft" number in the sense that it is, typically, quite uncertain and becomes more so as the sample size decreases. In real life we would much prefer to make categorical statements about the reliability of estimates than offer estimates of uncertain reliability.

---

‡Note that these X_i are discrete random variables.

§We mentioned in Chapter 1 that in many if not most practical problems, probabilities have to be estimated.

One of the central goals of parameter estimation is to construct events that are (nearly) certain to occur, that is, events whose probability is a "hard" number. That is not to say that soft numbers such as the estimated probability $\hat{p}$ are necessarily unreliable or useless. For example, suppose that rental-car dealership handling thousands of cars finds that at the end of year 1, n_E of its n_1 cars must be replaced due to wear and tear. Then, all things being equal, if the agency starts year 2 with n_2 cars, it could reasonably expect that approximately $\hat{p}n_2$ cars will have to be replaced at the end of year 2, perhaps a few more, perhaps a few less. Here $\hat{p} \triangleq n_E/n_1$ is the estimated probability that a car will have to be replaced by the end year 2. From the point of view of the executives of the company, the estimate $\hat{p}n_2$ is useful for year 2 planning and budgeting.

In Example 6.1-6 below we demonstrate how *firm* or *hard conclusions* can be drawn by applying basic principles of statistics.

Example 6.1-6

(*estimating the number of fish in a lake*) To illustrate how statistics can be used to generate meaningful certain events, consider the following problem. The United States Fish and Wildlife Services (FWS), a bureau of the Department of the Interior, is interested in estimating the percentage of freshwater bass in a large lake that for specificity we call Bass Lake. To that end, an "experiment" is performed where a net is used to capture a random sample of fish, which is subsequently examined for its bass content. In preparation for this experiment, we will denote the number of bass in the sample by n_B and the fixed sample size by n. Then we form the estimator $\hat{p} = n_B/n$, which is a random variable because n_B is a random variable[†]. We do not consider n a random variable because we can decide *a priori* how big a sample will be examined for its bass content. The true probability p that a fish pulled at random from the lake is a bass is the ratio of total bass in the lake to total fish in the lake; this number is unknown (and mildly variable over time since fish have a tendency to eat each other). At the risk of adding additional notation, we must carefully distinguish between the random variable n_B (a function) and its *realization*, which is a number. *Realizations*, whenever they don't add to confusion, will be superscripted with a prime. For example, a realization of n_B might yield $n'_B = 58$, $n = 133$ and the *estimated*[‡] probability that a fish selected at random will be a bass is $\hat{p}' = 58/133 = 0.44$. The range of the function n_B is the set of integers in the interval $[0, n]$. Of course, the realization $\hat{p}'$ is only a one-time estimate of the true probability p that a fish will be a bass and we would like to make a stronger statement about the number of bass in Bass Lake. Suppose we examine the fish in the sample one-by-one. Let

$$X_i \triangleq \begin{cases} 1, \text{ if the } ith \text{ fish is a bass,} \\ 0, \text{ else.} \end{cases}$$

then X_i is a Bernoulli RV with PMF $P_{X_i}(x) = p^x(1-p)^{1-x}$, for $x = 0, 1$ and zero else. The random sample $\{X_i, i = 1, \ldots, n\}$ consists of n i.i.d. observations on a generic random variable X, denoting whether a fish is a bass or not. We can think of X as a

[†]Here and a few other places we briefly depart from our use of capital letters to denote random variables.
[‡]The realization of an estimator is sometimes called an estimate, that is, a number.

population, that is, the fish population. The RV $Z \triangleq \sum_{i=1}^{n} X_i$ represents the total number of bass in a sample of n fish and $Z/n \triangleq \hat{p}$ is the estimator for p. Since Z is the sum of independent Bernoulli RVs it is a binomial R.V with PMF $b(k; n, p)$ (Example 4.8-1). Then Z has mean np and standard deviation $\sigma_Y = \sqrt{np(1-p)}$. We next create the (almost) certain event $E = \left\{np - 3\sqrt{np(1-p)} \le Z \le np + 3\sqrt{np(1-p)}\right\}$ and since Z is the sum of a large number of i.i.d. random variables we can use the Normal approximation to compute $P[E]$ as allowed by the Central Limit Theorem. Indeed this was done in Example 4.8-3, where $P[E]$ was computed to be 0.997. We can rewrite $P[E]$, using $Z/n \triangleq \hat{p}$, as $P[E] = P\left[(p-\hat{p})^2 \le \frac{9}{n}p(1-p)\right]$=0.997. We suggest that the reader verifies this result. The argument is a quadratic in p and solving for the roots p_1, p_2 of $(p-\hat{p})^2 = (9/n)\,p(1-p)$ will give the end points of the interval of integration about $\hat{p}$ that will yield an event probability of 0.997. These points are

$$p_1, p_2 = \frac{2\hat{p} + (9/n)}{2[1+(9/n)]} \mp \sqrt{\left(\frac{2\hat{p} + (9/n)}{2[1+(9/n)]}\right)^2 - \frac{\hat{p}^2}{1+(9/n)}} \quad . \tag{6.1-1}$$

For the numbers $n = 133, n_B = 58$, we get $\hat{p}' \approx 0.44$ and find that $\hat{p}_1' \approx 0.31, \hat{p}_2' \approx 0.57$. How do we interpret these results? First note that there is no probability associated with the *realized* interval [0.31, 0.57]; it either contains the true probability p or not; its length is 0.26 and $|\hat{p}' - p_1'| = |\hat{p}' - p_2'| \simeq 0.13$. The number $100 \times |\hat{p}' - p_1'|$ is sometimes called the *margin of error*, which in this case is 13 percent[†]. The interval with end points $[p_1, p_2]$ is a *random* interval because its end points are random variables; that is, they depend on the *estimate* $\hat{p}$. However, on the average, the interval will enclose the point p in 997 times in a thousand trials. We note that while the percentage of the bass in the lake is nowhere near zero or 100 percent, the probability that bass make up between 31 and 57 percent of the fish in Bass Lake is a near-certain event!

The above example illustrates how statistics has helped us to make a strong statement about the number of bass in Bass Lake. The statement might read like this: *Research has shown that 44 percent of the fish in Bass Lake are bass. The margin of error is ± 13 percent.*

Example 6.1-7

(*estimating dengue fever probability*) A newspaper article[‡] reported that inhabitants and visitors on the island of Key West in the State of Florida were being exposed to the virus that causes dengue fever. The illness is caused by the bite of a mosquito that carries the virus in its gut. While some in the island's tourist industry minimized the likelihood that a visitor would be infected with the virus, an independent study found that among 240 residents, presumably picked at random, 13 tested positive for the dengue fever virus. Some argued that the sample was too small to be accurate and that the dengue fever rate was much lower. Compute a 95 percent confidence interval on the true probability that a resident picked at random will test positive for dengue fever.

[†]It is not uncommon to describe the margin of error with an algebraic sign, for example in this case ±13%.

[‡]*The New York Times* of July 23, 2010.

Solution Our estimator for the true mean is $\hat{p} = K/n$ where K is a Binomial random variable i.e.

$$P_K[k \text{ successes in } n \text{ tries}] \triangleq b(k;n,p) = \binom{n}{k} p^k(1-p)^{n-k},$$

with $E[\hat{p}] = p$, and $\text{Var}[\hat{p}] = p(1-p)/n$. From the data we compute the mean estimate as $\hat{p}'$ =13/240=0.054. Since $n >> 1$, we use the Normal approximation to the Binomial and define the standard Normal random variable

$$X \triangleq (\hat{p}-p)/\sqrt{p(1-p)/n}$$

such that X:$N(0,1)$. Then a 95 percent confidence interval on p is found from solving $P(-x_{0.975} < X < x_{0.975}) = 2F_{SN}(x_{0.975}) - 1 = 0.95$ or $x_{0.975} \approx 1.96$. Then

$$P[-1.96 < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < 1.96] = 0.95.$$

Using the technique in Example 6.1-6, we find that the lower and upper limits of the 95 percent interval on p, in this case, are the roots of the polynomial $1.016p^2 - 0.124p + 0.003 = 0$, which are $p_l = 0.033, p_u = 0.089$. Thus we have a 95 percent confidence that the infection rate is from a low of 1 in 30 residents to a high of 1 in 11. Would this knowledge affect your plans to visit Key West?

6.2 ESTIMATORS

Estimators are functions of RVs that are used to estimate parameters but do not depend on the parameters themselves. We illustrate with some examples.

Example 6.2-1

(*truth in packaging*) A consumer protection agency (CPA) seeks to verify the information on the label of packages of cooked turkey breasts sold in supermarkets that says "70% meat, 30% water." The turkey breasts are produced by "Sundry Farms" and the CPA buys five "Sundry Farms" packages and checks for meat content percentage (mcp). With X_i denoting the mcp of the ith package, the CAP uses the function $\hat{\Theta}_1 = (1/n)\sum_{i=1}^{n} X_i$ to estimate the average mcp. It finds the following mcp's in the five packages ($n = 5$) *respectively*: 68, 82, 71, 65, 67 and obtains an average of 70.6 percent meat.

The 70.6 percent represents a *realization of the estimator* $\hat{\Theta}_1$ and is often called an *estimate*. If the CPA buys another set of five packages of cooked turkey breasts from "Sundry Farms," it would no doubt compute a slightly different estimate from the previous.

Example 6.2-2

(*truth in packaging continued*) The CPA seeks to estimate the variability in the meat content of "Sundry Farms"turkey breasts. It uses the formula $\hat{\Theta}_2 = \left((1/n)\sum_{i=1}^{n}\left(X_i - (1/n)\ \sum_{j=1}^{n} X_j\right)^2\right)^{1/2}$ with $n = 5$ and obtains approximately 6.0 percent meat variability using the data in the previous problem.

Example 6.2-3

(*truth in packaging continued*) The CPA is criticized for using $\hat{\Theta}_2$ in the previous problem as a measure of variability. It is suggested that the CPA use instead the estimator $\hat{\Theta}_3 = \left((1/(n-1))\sum_{i=1}^{n}\left(X_i - (1/n)\sum_{j=1}^{n} X_j\right)^2\right)^{1/2}$. Using $\hat{\Theta}_3$ with $n = 5$, the CPA computes a meat variability of 6.7 percent.

In what follows we shall find that $\hat{\Theta}_1$ is an unbiased and consistent estimator for the mean, $\hat{\Theta}_2$ is a biased, maximum likely estimator for the standard deviation, and $\hat{\Theta}_3$ is an unbiased and consistent estimator for the standard deviation. Other estimators are used to estimate Var $[X]$, the covariance matrix $\mathbf{K}$ and so on for the higher joint moments.

Some estimators have more desirable properties than others do. To evaluate estimators we introduce the following definitions.

Definition 6.2-1 An estimator† $\hat{\Theta}$ is a function of the observation vector $\mathrm{X} = (X_1, \ldots, X_n)^T$ that estimates θ but is not a function of θ. ■

Definition 6.2-2 An estimator $\hat{\Theta}$ for θ is said to be *unbiased* if and only if $E[\hat{\Theta}] = \theta$. The bias in estimating θ with $\hat{\Theta}$ is‡

$$|E[\hat{\Theta}] - \theta|. \quad ■$$

Definition 6.2-3 An estimator $\hat{\Theta}$ is said to be *a linear estimator* of θ if it is a linear function of the observation vector $\mathrm{X} \triangleq (X_1, \ldots, X_n)^T$, that is,

$$\hat{\Theta} = \mathbf{b}^T\mathbf{X}. \tag{6.2-1}$$

The vector $\mathbf{b}$ is an $n \times 1$ vector of coefficients that do not depend on $\mathbf{X}$. ■

Definition 6.2-4 Let $\hat{\Theta}_n$ be an estimator computed from n samples $X_1, ..., X_n$ for every $n \geq 1$. Then $\hat{\Theta}_n$ is said to be *consistent* if

$$\lim_{n\to\infty} P[|\hat{\Theta}_n - \theta| > \varepsilon] = 0. \qquad for\ every \qquad \varepsilon > 0. \tag{6.2-2}$$

The condition in Equation 6.2-2 is often referred to as *convergence in probability*. ■

Definition 6.2-5 An estimator $\hat{\Theta}$ is called *minimum-variance unbiased* if

$$E[(\hat{\Theta} - \theta)^2] \leq E[(\hat{\Theta}' - \theta)^2] \quad ■ \tag{6.2-3}$$

where $\hat{\Theta}'$ is any other estimator and $E[\hat{\Theta}'] = E[\hat{\Theta}] = \theta$.

Definition 6.2-6 An estimator $\hat{\Theta}$ is called a *minimum mean-square error* (MMSE) estimator if

$$E[(\hat{\Theta} - \theta)^2] \leq E[(\hat{\Theta}' - \theta)^2], \tag{6.2-4}$$

where $\hat{\Theta}'$ is any other estimator. ■

---

†The validity of estimating parameters as well as other objects, for instance probabilities, from repeated observations is based, fundamentally, on the law of large numbers and the Chebyshev inequality.

‡The bias is often defined without the magnitude sign. In that case the lines could be positive or negative.

There are several other properties of estimators that are deemed desirable such as efficiency, completeness, and invariance. These properties are discussed in books on statistics† and will not be discussed further here.

6.3 ESTIMATION OF THE MEAN

In Chapter 4 we showed that the *numerical average*, μ_s, *of a set of numbers* is the number that is simultaneously *closest* to all the numbers $x_1, x_2, \ldots, x_n$ in a set. In this sense μ_s can be regarded as the best representative of the set. Borrowing from mechanics, some think of the average as the *center of gravity* of the set. While the sample average doesn't tell the whole story, it is a useful *descriptor* for assessment in all sorts of situations. For example, if the average grade on a standardized test earned by students in School A is 92 and the average grade on the same test is 71 for students at School B, then, all other things being equal, one might conclude that School A does a better job of preparing its students than School B. If a large amount of data, suitably corrected for other factors (e.g., sex, income, race, lifestyle), showed that the average lifetime of smokers is 67 years while those of nonsmokers is 78 years, one could reasonably conclude that smoking is bad for your health.

Repeating Equation 4.1-1 here with a slight change of notation,

$$\mu_s(n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{6.3-1}$$

we observe that the *numerical average* depends on the *size* n of the number of the sample as well as the samples themselves. In our model we assume that the data are realizations of n i.i.d. observations on the generic random variable X; that is, x_1 is a one-time realization of the observation X_1, x_2 is a one-time realization of the observation X_2, and so forth. Each of the X_i is a function while x_i is a numerical value that the function obtains. We create the *mean-estimator function*

$$\hat{\mu}_X(n) \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \tag{6.3-2}$$

from the random sample $\{X_1, \ldots X_n\}$ to estimate the unknown parameter $\mu_X \triangleq E[X]$. We recognize that $\hat{\mu}_X(n)$ is the estimator $\hat{\Theta}_1$ introduced in Section 6.2. The object in Equation 6.3-2 is often called *the sample mean.* We use the hat to indicate that $\hat{\mu}_X(n)$ is an estimator and not the actual mean. Incidentally, it is useful to introduce at this point the *variance-estimator function* (VEF) or the *sample variance* as

$$\hat{\sigma}_X^2(n) \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2. \tag{6.3-3}$$

We recognize that VEF is the square of the estimator $\hat{\Theta}_3$ in Section 6.2. This is one of two VEFs that are in common use. The other one is

†See, for example,[6-1,Chapter 8].

$$\hat{\sigma}_X^2(n) \triangleq \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2. \tag{6.3-4}$$

We shall discuss the estimation of the variance in a later section but for now, we ask the reader's indulgence to take Equation 6.3-3 at face value. The estimation of σ_X^2 by the VEF in Equation 6.3-3 is, as we shall see later, an entirely reasonable thing to do. Among other attractive features we find that $E[\hat{\sigma}_X^2] = \sigma_X^2$, which is only asymptotically true for the VEF of Equation 6.3-4.

Properties of the Mean-Estimator Function (MEF)

The mean estimator given by Equation 6.3-2 is *unbiased* meaning that $E[\hat{\mu}_X(n) - \mu_X] = 0$. The proof of this important result is easy. We write

$$E[\hat{\mu}_X(n)] = E\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}E[X_i] = \frac{\mu_X}{n}\sum_{i=1}^{n}1 = \frac{\mu_X}{n}n = \mu_X. \tag{6.3-5}$$

An unbiased estimator is often, but not always, desired[†]. Another and important property of an estimator is that, in some sense, it gets better as we make more observations. For example, we would expect the MEF in Equation 6.3-2 to be more "reliable" if it is based on 100 rather than on 10 observations. One way to measure reliability is by way of the variance of the unbiased estimator. If the variance of the unbiased estimator is small, it is unlikely that a realization of $\hat{\mu}_X(n)$ will be very far from the true mean μ_X; if the variance is large, the realization might often be far from the true mean. Consider the variance of $\hat{\mu}_X(n)$. By definition this is

$$\begin{aligned}
\sigma_{\hat{\mu}}^2(n) &\triangleq E\left[(\hat{\mu}_X(n) - (E\left[\hat{\mu}_X(n)\right]))^2\right] = E\left[(\hat{\mu}_X(n) - \mu_X)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)\right)^2\right] \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \mu_X)^2\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(X_i - \mu_X)(X_j - \mu_X)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}E[(X_i - \mu_X)^2] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n}E\left[(X_i - \mu_X)(X_j - \mu_X)\right] \\
&= \sigma_X^2/n.
\end{aligned} \tag{6.3-6}$$

In line 1, the term on the right uses the unbiasedness of the MEF. Line 2 uses the definition of the MEF and multiplies and divides $\mu_X$by n. Line 3 uses that the square of a sum is

[†]One may tolerate a small amount of bias if the estimator has other desirable properties.

the sum of squares plus the sum of cross-term products with nonequal indexes. Line 4 uses the linearity of the expectation operator, and line 5 takes advantage of the fact that for $i \neq j, X_i$ andX_j are independent and therefore $E[(X_i - \mu_X)(X_j - \mu_X)] = 0$. At this point we invoke the Chebyshev inequality (see Section 4.4) and apply it to $\hat{\mu}_X(n)$. Then for any $\delta > 0$

$$P[|\hat{\mu}_X(n) - \mu_X| \geq \delta] \leq \frac{\text{Var}(\hat{\mu}_X(n))}{\delta^2} = \frac{\sigma_X^2}{n\delta^2} \overset{n\to\infty}{\longrightarrow} 0 \quad . \tag{6.3-7}$$

Equations 6.3-6 and 6.3-7 are among the most important results in all of statistics. Equation 6.3-6 says that the variance of the mean estimator *decreases with increasing* n and hence can be made arbitrarily small by choosing a large enough sample size. Specifically, the variance of the mean estimator is numerically equal to the variance of the observation variable divided by the sample size. This is true so long as the observation variable X *has finite variance.* Equation 6.3-7 says that the event that the absolute deviation between the true mean and the MEF exceeds a certain value—no matter how small that value is—becomes highly improbable when the sample size is made large enough. An estimator that obeys Equations 6.3-6 and 6.3-7 is said to be *consistent.*

Example 6.3-1

(*effect of sample size on estimating the mean*) We wish to compute $P[|\hat{\mu}_X(n) - \mu_X| \leq 0.1]$ when X is Normal with $\sigma_X = 3$. To illustrate the effect of sample size we use two random samples: a small sample ($n = 64$) and a large sample ($n = 3600$). We write

$$\begin{aligned}
&P[-0.1 < \hat{\mu}_X(n) - \mu_X < 0.1] \\
&= P[-0.1\sqrt{n}/\sigma_X < Y < 0.1\sqrt{n}/\sigma_X] \\
&= 2\,\text{erf}\left(\frac{0.1\sqrt{n}}{\sigma_X}\right) \\
&= 2\,\text{erf}(0.0333\sqrt{n}),
\end{aligned}$$

where $Y \triangleq (\hat{\mu}_X - \mu_X)/(\sigma_X/\sqrt{n})$ is distributed as $N(0,1)$. When $n = 64$, $P[|\hat{\mu}_X(n) - \mu_X| \leq 0.1] \approx 0.2$. We can interpret this result as saying that in a thousand trials involving sample sizes of 64, in only about 200 outcomes will the mean estimate deviate from the true mean by 0.1 or less. For $n = 3600$, we compute $P[|\hat{\mu}_X(n) - \mu_X| \leq 0.1] \simeq 0.95$, which implies that the event $\{|\hat{\mu}_X(n) - \mu_X| \leq 0.1\}$ will occur in about 950 out of a 1000 trials. The implication is that in a single trial, the event $\{|\hat{\mu}_X(n) - \mu_X| \leq 0.1\}$ will almost certainly happen when $n = 3600$.

Example 6.3-2

(*how many samples do we need to get a 95 percent confidence interval on the mean?*) We want to compute a 95 percent *confidence interval* on the mean of a Normal random variable X. How many observations $X_1, \ldots, X_n$ on X do we need? More to the point, what parameters determinate the length and location of the interval? The terminology "95 percent confidence interval" merely means that we seek the end points of the shortest (or near-shortest) interval on the real line such that we expect that in 950 or so cases out of 1000 the interval will enclose the true mean. In terms of a probability we write

$$P[|\hat{\mu}_X(n) - \mu_X| \leq \gamma_{0.95}] = 0.95, \tag{6.3-8}$$

where the number $\gamma_{0.95}$ is a number to be determined and its subscript reminds us that it is a 95 percent confidence interval we seek. We recall that $\hat{\mu}_X(n)$ is $N(\mu_X, \sigma_X^2/n)$ so that

$$Y \triangleq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}} \tag{6.3-9}$$

is $N(0,1)$. Then, rewriting Equation 6.3-8 with Y in mind, we obtain

$$\begin{aligned} 0.95 &= P[-\gamma_{0.95} \leq \hat{\mu}_X(n) - \mu_X \leq \gamma_{0.95}] \\ &= P\left[-\gamma_{0.95}\sqrt{n}/\sigma_X \leq Y \leq \gamma_{0.95}\sqrt{n}/\sigma_X\right]. \\ &= 2F_{SN}(\gamma_{0.95}\sqrt{n}/\sigma_X) - 1 \end{aligned} \tag{6.3-10}$$

In line 2 we converted the RV $\hat{\mu}_X(n) - \mu_X$ into an N (0, 1) random variable Y. In line 3 we expressed this probability in terms of the standard Normal CDF. The last line of Equation 6.3-10 yields the result we seek, that is $F_{SN}(\gamma_{0.95}\sqrt{n}/\sigma_X)$= 0.975. As on other occasions we use the symbol $F_{SN}(z_u) = u$ to denote the *standard Normal* (*SN*) CDF. The number z_u is called the u-percentile of the standard Normal. From the tables of the CDF (see Appendix G) we find that $z_{0.975} = 1.96$. But since $z_{0.975} = \gamma_{0.95}\sqrt{n}/\sigma_X$, we deduce that $\gamma_{0.95}\sqrt{n}/\sigma_X = 1.96$ or, equivalently, $\gamma_{0.95} = 1.96\sigma_X/\sqrt{n}$. Returning to the problem at hand, we note that the event $\{|\hat{\mu}_X(n) - \mu_X| \leq \gamma_{0.95}\}$ is the same as the event $\{\hat{\mu}_X(n) - \gamma_{0.95} \leq \mu_X \leq +\hat{\mu}_X(n) + \gamma_{0.95}\}$. Then, from the middle line of Equation 6.3-10 we get that (on the average) a *shortest* 95 percent confidence interval for μ_X as

$$\left[-1.96\frac{\sigma_X}{\sqrt{n}} + \hat{\mu}_X(n),\ 1.96\frac{\sigma_X}{\sqrt{n}} + \hat{\mu}_X(n)\right]. \tag{6.3-11}$$

Of course this result can be generalized to other than 95 percent confidence intervals. Suppose we seek a δ-confidence interval (here we specified δ= 0.95). Then a δ-confidence interval on μ_X is

$$\left[-z_{(1+\delta)/2}\frac{\sigma_X}{\sqrt{n}} + \hat{\mu}_X(n), z_{(1+\delta)/2}\frac{\sigma_X}{\sqrt{n}} + \hat{\mu}_X(n)\right]. \tag{6.3-12}$$

How do we know that, on the average, it is the shortest interval? Because of the symmetry of the Normal pdf, the largest amount of probability mass is at the center. Any other 95 percent interval will require more *support*, that is, need a longer length.

Let us return to what was asked for. The question as to how many samples are needed for a shortest 95 percent confidence interval cannot be determined if σ_X is not known. Clearly, by choosing a large enough interval, for example, a ten-sigma width on either side of $\hat{\mu}_X(n)$, we shall get a 95 percent (and more!) confidence even when the number of samples, n, is small. But with a ten-sigma width on either side the interval *will not be the shortest and will prove useless because it is too large.* So let us assume that it is the *shortest* interval that we seek. Then the interval will be centered about $\hat{\mu}_X(n)$ and have width $W_{0.95} = 2 \times 1.96\sigma_X/\sqrt{n}$. So clearly, the ratio $\sigma_X/\sqrt{n}$ determines the width of the interval.

If σ_X is known (this is unlikely in practice), then we can determine how many samples we need to obtain a confidence interval of a specified length. For an arbitrary δ-confidence interval, the width of the confidence interval is

$$W_\delta = 2 \times z_{(1+\delta)/2} \times \sigma_X/\sqrt{n}. \tag{6.3-13}$$

Not surprisingly we find from Equation 6.3-13 that the interval gets wider (which increases our uncertainty as to the true mean) when the standard deviation of X increases but gets smaller (which decreases our uncertainty as to the true mean) when the number of samples increases. Also the interval gets wider when the demanded percent confidence increases. Does this make sense?

Procedure for Getting a δ-confidence Interval on the Mean of a Normal Random Variable When σ_X Is Known

(1) Choose a value of δ and compute $(1+\delta)/2$;
(2) From the tables of the CDF for the standard Normal find the percentile $z_{(1+\delta)/2}$ such that $F_{SN}(z_{(1+\delta)/2}) = (1+\delta)/2$;
(3) Obtain the realizations of X_i, $i = 1, \ldots, n$. Label these numbers x_i, $i = 1, \ldots, n$. Compute the numerical average $\mu_s = \frac{1}{n}\sum_{i=1}^{n} x_i$;
(4) Compute the interval $\left[-z_{(1+\delta)/2}\frac{\sigma_X}{\sqrt{n}} + \mu_s, z_{(1+\delta)/2}\frac{\sigma_X}{\sqrt{n}} + \mu_s\right]$.

Up until now, we have assumed that σ_X is known. However, σ_X is typically not known (Can you think of a situation where we do not know μ_X but know σ_X?) One possible solution to this problem is to replace σ_X in Equation 6.3-11 by an estimated value of it, for example, $\hat{\sigma}_X(n)$, the square root of Equation 6.3-3, and continue with our assumption that Y is Normal. But in fact Y would not be Normal because of the randomness in $\hat{\sigma}_X(n)$ and this might not yield accurate results especially when the sample size is not large. Not knowing σ_X requires that we seek another approach for determining a prescribed confidence interval. Such an approach is furnished by the t-distribution discussed below.

Confidence Interval for the Mean of a Normal Distribution When σ_X Is Not Known

In general, the distributions one encounters in *statistics* are often of an algebraic form that is more complex than those we encounter in elementary probability. One of these is the so-called "student's" t-distribution introduced by W. S. Gossett in connection with his work of computing a confidence interval for the mean of a Normal distribution when the *variance is not known.* Gossett is considered one of the founders of modern *statistics* but is better known by his pen name *Student*†. As we saw in our previous discussion, the problem of finding the end points of a confidence interval involves the distribution of the $N(0,1)$ RV

†1876–1937. Much secrecy enveloped his work on statistical quality control at the Dublin brewery of Arthur Guinness & Son. For this reason he used the pseudo name "Student."

$$Y \triangleq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}}.$$

However, without knowledge of σ_X we cannot find the end-points that define the confidence interval. So we create a new RV by replacing σ_X by

$$\hat{\sigma}_X(n) = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2\right)^{1/2},$$

which is merely the sigma value derived from the VEF of Equation 6.3-3. This new RV is defined by

$$T_{n-1} \triangleq \frac{\hat{\mu}_X(n) - \mu_X}{\left(\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2\right)^{1/2}} = \frac{\hat{\mu}_X(n) - \mu_X}{\hat{\sigma}_X(n)/\sqrt{n}} \tag{6.3-14}$$

and is said to have a t-distribution with $n-1$ degrees of freedom for $n = 2, 3\ldots$ We do not treat T_{n-1} as an approximation to a standard Normal RV. As n changes, we generate a family of t-distributions. We denote the pdf associated with T_{n-1} by $f_T(x; n-1)$. The important thing to observe is that T_{n-1} does not involve the unknown σ_X, a fact that enables us to compute confidence intervals on the mean μ_X, something we could not do using the RV in Equation 6.3-9.

It is important for the reader to understand that in creating the t-distribution we did *not approximate* σ_X by $\hat{\sigma}_X$. The brilliance of the contribution of Gossett was in avoiding approximations required to use the Normal distribution and working instead with T_{n-1} and its distribution.

For insight, we can rewrite Equation 6.3-14 as

$$T_{n-1} \triangleq \frac{(\hat{\mu}_X(n) - \mu_X)\sqrt{n}/\sigma_X}{\left(\frac{1}{(n-1)}\sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}_X(n)}{\sigma_X}\right)^2\right)^{1/2}} = \frac{Y}{(Z_{n-1}/n-1)^{1/2}}, \tag{6.3-15}$$

where $Y \triangleq (\hat{\mu}_X(n) - \mu_X)\sqrt{n}/\sigma_X : N(0,1)$ and $Z_{n-1} \triangleq \sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}_X}{\sigma_X}\right)^2$ has a χ^2_{n-1} pdf with $n-1$ degrees of freedom. When spelled out the symbol χ^2 is written Chi-square (pronounced *ky-square* as in *sky-square*). The subscript of the Chi-square RV gives the number of degrees of freedom (DOF) and the RV range is $(0, \infty)$. This implies that the CDF $F_{\chi^2}(z; n) = 0$ for $z \leq 0$ for every integer $n \geq 1$. The χ^2 distribution was introduced in Chapter 2 and is sometimes called a *sampling distribution* because it involves i.i.d. samples of a population X. It is not obvious but Y and Z_{n-1}, although sharing the same $X_i, i = 1, \ldots, n$, can be shown to be statistically independent (see Appendix G). From Equation 6.3-14 we see that that the t-random variable is the ratio of a standard Normal RV (numerator) to the square root of a quotient of a Chi-square RV divided by the DOF.

For large values of n the t-distribution will not be that different from the Normal (see Figure 6.3-1). Indeed the pdf of T_{n-1} is centered at the origin and symmetrical about it. In

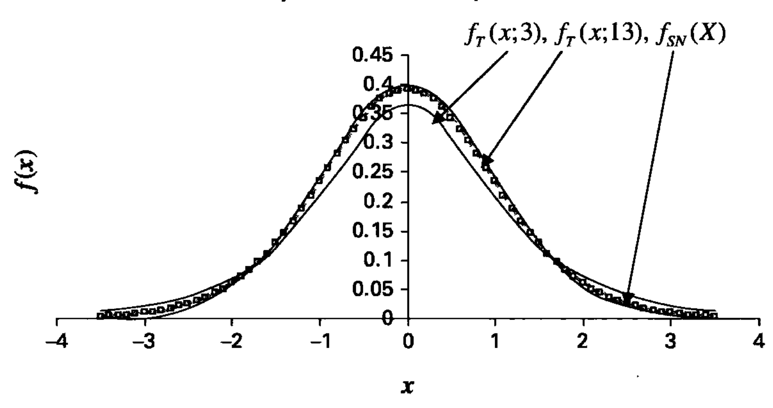


Figure 6.3-1 The probability density function of the T random variable has a shape similar to that of the Normal pdf, especially as the number of degrees-of-freedom get larger. Here is shown the t-pdf for $n = 3$ (peaks at 0.36); $n = 13$ (the curve with the boxes that peak at 0.39); and the *SN* pdf. Except for a barely observed variation in the tails, the $n = 13$ t-distribution is virtually identical with the Normal.

seeking the shortest confidence interval for μ_X, we consider the event $\{-t_{\delta/2} \leq T_{n-1} \leq t_{\delta/2}\}$. The probability of this event is

$$P[-t_{(1+\delta)/2} \leq T_{n-1} \leq t_{(1+\delta)/2}] = \delta, \tag{6.3-16}$$

where, as before, $100 \times \delta$ is the *assigned percent confidence* for interval on μ_X. With the CDF for the T_{n-1} RV denoted by $F_T(t; n-1) = \int_{-\infty}^{t} f_T(x; n-1)dx$, we find that

$$\delta = 2F_T(t_{(1+\delta)/2}, n-1) - 1$$

or, equivalently,

$$F_T(t_{(1+\delta)/2}; n-1) = \frac{1+\delta}{2}. \tag{6.3-17}$$

From the tables of the cumulative t-distribution with DOF $n-1$ in Appendix G, we can determine *the t-percentile* $t_{(1+\delta)/2}$. Finally, from Equations 6.3-14 and 6.3-16, we obtain

$$P\left[\hat{\mu}_X(n) - \frac{t_{(1+\delta)2}\hat{\sigma}_X(n)}{\sqrt{n}} \leq \mu_X \leq \hat{\mu}_X(n) + \frac{t_{(1+\delta)/2}\hat{\sigma}_X(n)}{\sqrt{n}}\right] = \delta,$$

which gives as a 100δ percentage confidence interval

$$\left[\hat{\mu}_X(n) - \frac{t_{(1+\delta)/2}\hat{\sigma}_X(n)}{\sqrt{n}}, \hat{\mu}_X(n) + \frac{t_{(1+\delta)/2}\hat{\sigma}_X(n)}{\sqrt{n}}\right]. \tag{6.3-18}$$

The width of the confidence interval is

$$W_\delta = 2\frac{t_{(1+\delta)/2}\hat{\sigma}_X(n)}{\sqrt{n}}. \tag{6.3-19}$$

Procedure for Getting a δ-Confidence Interval Based on n Observations on the Mean of a Normal Random Variable when σ_X Is Not Known

(1) Choose a value of δ and compute $(1+\delta)/2$;
(2) From the tables of the CDF for T_{n-1}, find the t-percentile number $t_{(1+\delta)/2}$ such that $F_T(t_{(1+\delta)/2}; n-1) = (1+\delta)/2$;
(3) Obtain the realizations of $X_i, i = 1, \ldots, n$. Label these numbers $x_i, i = 1, \ldots, n$. Compute the realizations of $\hat{\mu}_X(n), \hat{\sigma}_X(n)$;
(4) Compute the numerical realization of the interval $\left[\hat{\mu}_X(n) - \frac{t_{(1+\delta)/2}\hat{\sigma}_X(n)}{\sqrt{n}},\right.$ $\left.\hat{\mu}_X(n) + \frac{t_{(1+\delta)/2}\hat{\sigma}_X(n)}{\sqrt{n}}\right]$.

Example 6.3-3

(*confidence interval on μ_X when σ_X is unknown-Normal case*) Twenty-one i.i.d. observations ($n = 21$) are made on a Gaussian RV X. These observations are denoted as $X_1, X_2, \ldots, X_{21}$. Based on the data, the realizations of $\hat{\mu}_X(n)$ and $\hat{\sigma}_X(n)/\sqrt{n}$ are, respectively, 3.5 and 0.45. A 90 percent confidence interval on $\hat{\mu}_X(n)$ is desired.

Solution Since $P[-t_{0.95} \leq T_{20} \leq t_{0.95}] = 0.9$, we obtain from Equation 6.3-17 $F_T(t_{0.95}, 20) = 0.5(1+0.9) = 0.95$. Entering the student-t tables at $F = 0.95$ and $n = 20$ we obtain $t_{0.95} = 1.725$. The corresponding interval, from Equation 6.3-18, is $[3.5 - 1.725 \times 0.45, 3.5 + 1.725 \times 0.45] = [2.72, 4.28]$. The width of the interval is $W_\delta \approx 2 \times 1.725 \times 0.45 = 1.55$:

Interpretation of the Confidence Interval

The confidence interval generated from a series of realizations either will or will not include the true mean of X, which is a number unknown to us. Therefore, what does it mean to say that we have a "90 percent" confidence interval? The answer to this question goes to the heart of the meaning of probability, namely the frequency of a desirable outcome in repeated trials. Put succinctly, a "90 percent" confidence interval means that, say, in a thousand trials, one will observe that the interval covers the true mean about 900 times. Will we observe exactly 900 true-mean coverage? Not likely, but a success rate of 900 is the most likely outcome.

6.4 ESTIMATION OF THE VARIANCE AND COVARIANCE

We make n observations $X_1, X_2, \ldots, X_n$ on a Normal RV X with mean μ_X and variance σ_X^2. If μ_X is known then an unbiased VEF is computed from the random sample as

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2 \tag{6.4-1}$$

and it is not difficult to show that $\hat{\sigma}_X^2(n)$ is an unbiased, consistent estimator of σ_X^2. If the mean is not known, then the VEF

$$\hat{\sigma}_X^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \tag{6.4-2}$$

is an unbiased and consistent estimator of σ_X^2 .

Unbiasedness of $\hat{\sigma}_X^2(n)$ of Equation 6.4-2. Consider

$$E\left[\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right]$$

$$= E\left[\sum_{i=1}^{n}\left\{X_i^2 - \frac{2}{n}X_i^2 - \frac{2}{n}\sum_{\substack{j=1\\ j\neq i}}^{n}X_iX_j + \frac{1}{n^2}\sum_{j=1}^{n}X_j^2 + \frac{2}{n^2}\sum_{k=1}^{n}\sum_{j>k}^{n}X_jX_k\right\}\right]$$

$$= (n-1)\sigma^2. \tag{6.4-3}$$

In obtaining Equation 6.4-3, we used the fact that $E[X_i^2] = \sigma^2 + \mu^2, i = 1, \ldots, n$. Clearly if

$$E\left[\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right] = (n-1)\sigma^2,$$

then

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right] = \sigma^2. \tag{6.4-4}$$

But the quantity inside the square brackets is $\hat{\sigma}_X^2(n)$ of Equation 6.4-2. Hence $\hat{\sigma}_X^2(n)$ is unbiased for σ^2.

Consistency of $\hat{\sigma}_X^2(n)$ of Equation 6.4-2 The variance of $\hat{\sigma}_X^2(n)$ is given by $\text{Var}[\hat{\sigma}_X^2(n)] = E[(\hat{\sigma}_X^2(n) - \sigma^2)^2]$

$$= E\left[\frac{1}{(n-1)^2}\left\{\sum_{i=1}^{n}(X_i - \hat{\mu})^4 + \sum_{i\neq j}^{n}\sum^{n}(X_i - \hat{\mu})^2(X_j - \hat{\mu})^2\right\}\right.$$

$$\left. +\sigma^4 - \frac{2\sigma^2}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right].$$

A straightforward calculation shows that for $n >> 1$

$$\text{Var}[\hat{\sigma}_X^2(n)] \simeq \frac{1}{n}c_4, \tag{6.4-5}$$

where $c_4 \triangleq E[(X_1 - \mu)^4]$ (see Equation 4.3-2a). Assuming that c_4 (the fourth-order central moment) exists, we once again use the Chebyshev inequality to write that

$$P[|\hat{\sigma}_X^2(n) - \sigma^2| > \varepsilon] \leq \frac{\text{var}[\hat{\sigma}_x^2(n)]}{\varepsilon^2} \simeq \frac{c_4}{n\varepsilon^2} \xrightarrow{n\to\infty} 0. \tag{6.4-6}$$

Hence $\hat{\sigma}_X^2(n)$ is a consistent estimator for σ^2.

Example 6.4-1

(*computing the numerical sample mean and numerical sample variance of a Normal random variable*) Ten observations are made on a Normal RV $X{:}N(3, 1/10)$. The realizations are: 3.12, 2.87, 3.04, 2.77, 2.89, 3.34, 3.51, 2.44, 3.28, and 2.95. To compute the numerical sample mean and the numerical sample variance, we proceed as follows:

The numerical sample mean is computed as

$$\mu_s = \frac{1}{10}(3.12 + 2.87 + 3.04 + 2.77 + 2.89 + 3.34 + 3.51 + 2.44 + 3.28 + 2.95) = 3.02$$

The numerical sample variance is computed as

$$\begin{aligned}\sigma_s^2 &= \frac{1}{9}(0.01 + 0.225 + 0.0004 + 0.0625 + 0.0169 + 0.1024 + 0.2401 \\ &\quad + 0.3364 + 0.0676 + 0.0049) \\ &= 0.096.\end{aligned}$$

In signal processing the ratio $(\mu_s/\sigma_s)^2$ is sometimes called the ***signal-to-noise (power) ratio***; in this case it is 95. It is commonly given in decibels (dB), which in this case is $10 \times \log_{10} 95 = 19.8$ dB.

Confidence Interval for the Variance of a Normal Random variable

Determining a confidence interval for the variance involves the χ^2 distribution. Suppose we make n i.i.d. observations on the Normal RV X and label these observations as X_1, $X_2, \ldots, X_n$. Then, for each i

$$U_i \triangleq \frac{X_i - \mu_X}{\sigma_X} \tag{6.4-7}$$

is $N(0,1)$ and $Z_n \triangleq \sum_{i=1}^{n} U_i^2$ is Chi-square distributed with a DOF of n. The χ^2 pdf is shown in Figure 6.4-1 and is denoted by $f_{\chi^2}(x; n)$. If μ_X is not known in Equation 6.4-7 and we replace it with $\hat{\mu}_X(n)$ from Equation 6.3-2 we create a new RV

$$V_i \triangleq \frac{X_i - \hat{\mu}_X(n)}{\sigma_X} \tag{6.4-8}$$

and the sum $Z_{n-1} = \sum_{i=1}^{n} V_i^2$ is also Chi-square but with $n - 1$ degrees of freedom.

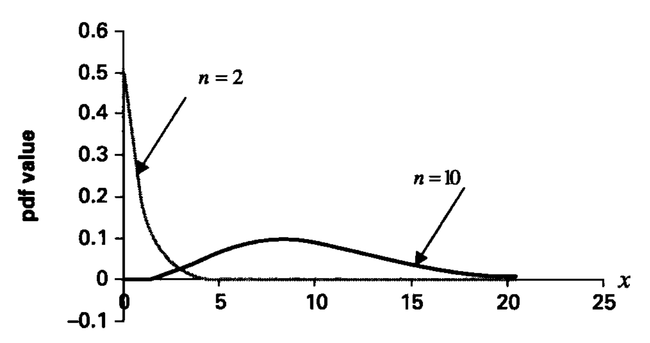


Figure 6.4-1 The Chi-square pdf for $n = 2$ (curve with value 0.5 at the origin) and $n = 10$. For all values of $n > 2$, the pdf will be zero at the origin.

Example 6.4-2

(*computing the degrees of freedom of a Chi-square RV*) With the V_i defined in Equation 6.4-8, the random variable $\sum_{i=1}^{2} V_i^2$ is Chi-square with a DOF of unity. We can see this with the help of a little bit of algebra. We find that $V_1^2 + V_2^2 = \left[(X_1 - X_2)/\sigma_X\sqrt{2}\right]^2$. But $U \triangleq (X_1 - X_2)/\sigma_X\sqrt{2}$ is $N(0,1)$ and hence in the sum $Z_n \triangleq \sum_{i=1}^{n} U_i^2$ there is only one nonzero term, that is, $U^2 = Z_1$.

To find a confidence interval on σ_X^2 at level, say, δ (e.g., $\delta = 0.95$, $\delta = 0.98$, $\delta = 0.99$), we begin with

$$W_{n-1} \triangleq \sum_{i=1}^{n} V_i^2 = \frac{1}{\sigma_X^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2$$

and seek numbers a, b such that $P[a \le W_{n-1} \le b] = P\left[a \le \frac{1}{\sigma_X^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \le b\right] = \delta$. For $a > 0$, $b > 0$, and $b > a$ the event $\{\zeta : a \le \frac{1}{\sigma_X^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \le b\}$ is identical with the event $\{\zeta : \frac{1}{b}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \le \sigma_X^2 \le \frac{1}{a}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2\}$. Hence the width of the confidence interval for the variance is†

$$W_\delta(a,b) = (\frac{1}{a} - \frac{1}{b})\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2. \tag{6.4-9}$$

Since $W_{n-1} = \frac{1}{\sigma_X^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2$ is χ_{n-1}^2 we solve for the numbers a, b from $P[a \le W_{n-1} \le b] = F_{\chi^2}(b; n-1) - F_{\chi^2}(a; n-1)$. To avoid the algebraic difficulties associated with

†Please do not confuse the width symbol $W_\delta(a,b)$ with the χ^2 random variable symbol W_n.

finding the *shortest* interval, we find numbers a, b that give a near-shortest interval as follows: The probability that W_{n-1} lies *outside* the interval is $1-\delta = 1-F_{\chi^2}(b; n-1)+F_{\chi^2}(a; n-1)$; if we denote $1-\delta$ as the "error probability" and assign $1-F_{\chi^2}(b; n-1) = (1-\delta)/2$ and $F_{\chi^2}(a; n-1) = (1-\delta)/2$, then we have divided the overall "error probability" into equal area-halves under the tails of the χ^2_{n-1} pdf. It then follows that $a = x_{(1-\delta)/2}$, that is, $F_{\chi^2}(x_{(1-\delta)/2}; n-1) = (1-\delta)/2$ and $b = x_{(1+\delta)/2}$, that is, $F_{\chi^2}(x_{(1+\delta)/2}; n-1) = (1+\delta)/2$. The numbers $x_{(1-\delta)/2}$ and $x_{(1+\delta)/2}$ are called, respectively, the $(1-\delta)/2$ and $(1+\delta)/2$ percentiles of the χ^2_{n-1} RV. The δ-confidence interval for the variance is

$$\left\{\frac{1}{x_{(1+\delta)/2}}\sum_{i=1}^{n}(X_i-\hat{\mu}_X(n))^2, \frac{1}{x_{(1-\delta)/2}}\sum_{i=1}^{n}(X_i-\hat{\mu}_X(n))^2\right\}$$

and its length L is

$$\left\{\left(\frac{1}{x_{(1-\delta)/2}}-\frac{1}{x_{(1+\delta)/2}}\right)\sum_{i=1}^{n}(X_i-\hat{\mu}_X(n))^2\right\}.$$

Example 6.4-3

Sixteen i.i.d. observations are made on $X{:}N(\mu_X, \sigma_X^2)$. A confidence interval on σ_X^2 is required. Find the numbers a, b that will give a near-shortest 95 percent confidence interval σ_X^2 using the "equal error probability" rule.

Solution $F_{\chi^2}(a; 15) = F_{\chi^2}(x_{0.025}; 15) = 0.025$. $F_{\chi^2}(x_{0.975}; 15) = 0.975$. From the table of the Chi-square distribution, we find $a = x_{0.025} = 6.26$ and $b = x_{0.925} = 27.5$.

Estimating the Standard Deviation Directly

We can estimate the standard deviation σ_X from

$$\hat{\sigma}_X(n) = \left(\frac{1}{n-1}\sum_{i=1}^{n}[X_i-\hat{\mu}_X(n)]^2\right)^{1/2} \tag{6.4-10}$$

but this involves computing $\hat{\sigma}_X^2(n)$ first. Another approach estimates σ_X directly. Consider two i.i.d. observations X_1, X_2 on the generic RV X. Let $Z \triangleq \max(X_1, X_2)$, $\hat{\mu} \triangleq (X_1+X_2)/2$. The pdf of Z is readily computed as $f_Z(z) = 2F_X(z)f_X(z)$, where $F_X(z)$ and $f_X(z)$ are, respectively, the CDF and pdf of X. Now consider the estimator $\hat{\sigma}_X$

$$\hat{\sigma}_X \triangleq \sqrt{\pi}(Z-\hat{\mu}_X) \tag{6.4-11}$$

and compute $E[\hat{\sigma}_X] \triangleq \sqrt{\pi}E[(Z-\hat{\mu}_X)] = \sqrt{\pi}(E[Z]-\mu_X)$. The computation of $E[Z]$ when X is Normal can be done with the aid of standard tables of integrals (see *Handbook of Mathematical Functions*, M. Abramowitz and I. A. Stegun, eds., Dover, New York, 1970,

p. 303, formula 7.4.14), or with Maple, MathCAD, Mathematica, etc. We find that $E[Z] = \mu_X + \frac{1}{\sqrt{\pi}}\sigma_X$ so that $E[\hat{\sigma}_X] = \sigma_X$. Hence $\hat{\sigma}_X \triangleq \sqrt{\pi}(Z - \hat{\mu}_X)$ *is an unbiased estimator for* σ_X.

Example 6.4-4

(*one-shot estimation of* σ_X) Two realizations of $X : N(\mu_X, \sigma_X^2)$ are obtained as 3.8, 4.1. Then (primes indicate realizations) $Z' = \max(3.8, 4.1) = 4.1$, $\hat{\mu}' = (3.8 + 4.1)/2 = 3.95$, and $\hat{\sigma}'_X = 0.26$. Computing $\hat{\sigma}'_X$ from Equation 6.3-6 yields 0.21.

To compute the variance of the *standard deviation estimator function* (SDEF) in Equation 6.4-11 we write:

$$\mathrm{Var}(\hat{\sigma}_X) = \pi(E[Z^2] + E[\hat{\mu}_X^2] - 2E[Z\hat{\mu}_X]) - \sigma_X^2.$$

This computation takes some work but the result is

$$\mathrm{Var}(\hat{\sigma}_X) = \left(\frac{\pi}{2} - 1\right)\sigma_X^2 \simeq 0.57\sigma_X^2. \tag{6.4-12}$$

In practice we would not want to estimate σ_X from only two observations on X. Suppose we make n (even) observations on X, which we denote as $X_1, X_2, \ldots, X_n$ and pair them as $\{X_1, X_2\}, \ldots, \{X_{n-1}, X_n\}$. Let

$$\hat{\sigma}_X^{(1)} \triangleq \sqrt{\pi}\,(\max(X_1, X_2) - 0.5(X_1 + X_2))$$
$$\hat{\sigma}_X^{(2)} \triangleq \sqrt{\pi}\,(\max(X_3, X_4) - 0.5(X_3 + X_4))$$
$$\vdots$$
$$\hat{\sigma}_X^{(n/2)} \triangleq \sqrt{\pi}\,(\max(X_{n-1}, X_n) - 0.5(X_{n-1} + X_n))$$

and define

$$\hat{\sigma}_{ave} \triangleq \frac{1}{n/2}\sum_{i=1}^{n/2}\hat{\sigma}_X^{(i)}. \tag{6.4-13}$$

Then Var $(\hat{\sigma}_{ave}) \triangleq \frac{4}{n^2}\sum_{i=1}^{n/2}\mathrm{Var}(\hat{\sigma}_X)$, which gives

$$\mathrm{Var}(\hat{\sigma}_{ave}) \approx \frac{1.04}{n}\sigma_X^2. \tag{6.4-14}$$

It is straightforward to show that $\hat{\sigma}_{ave}$ is a consistent estimator for σ_X; we leave this as an exercise for the reader. A confidence interval for σ_X based on estimating σ_X with $\hat{\sigma}_X \triangleq \sqrt{\pi}(Z - \hat{\mu})$ is discussed in [6-3] and [6-4].

Estimating the covariance

The *covariance*, defined by

$$c_{11} \triangleq \mathrm{Cov}[XY] = E\left[(X - \mu_X)(Y - \mu_Y)\right], \tag{6.4-15}$$

is classically estimated from the *covariance estimating function* (CEF)

$$\hat{c}_{11} \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_X(n)) \times (Y_i - \hat{\mu}_Y(n)), \tag{6.4-16}$$

where $\{X_i, Y_i, i = 1, \ldots, n\}$ are n *paired* i.i.d. observations. We leave it to the reader to show that $\hat{c}_{11}$ is an unbiased and consistent estimator for c_{11}. The *normalized covariance*, also called the *correlation coefficient*, is defined as

$$\rho_{XY} \triangleq \frac{c_{11}}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{6.4-17}$$

It is estimated from

$$\hat{\rho}_{XY} \triangleq \frac{\hat{c}_{11}}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}} = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu}_X(n)) \times (Y_i - \hat{\mu}_Y(n))}{\left(\sum_{i=1}^{n} (X_i - \hat{\mu}_X(n))^2 \sum_{i=1}^{n} (Y_i - \hat{\mu}_Y(n))^2\right)^{1/2}}. \tag{6.4-18}$$

The distribution of $\hat{\rho}_{XY}$ is not available in closed form. However, a confidence interval for ρ_{XY} can be found using more advanced methods [6-1].

6.5 SIMULTANEOUS ESTIMATION OF MEAN AND VARIANCE

If we seek, say, a 95 percent confidence region on both μ_X and σ_X^2 we take advantage of the RVs $\hat{\mu}_X(n)$ and $\hat{\sigma}_X^2(n)$ being independent. Thus, we may write

$$P\left[-a \leq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}} \leq a, b \leq \frac{1}{\sigma_X^2} \sum_{i=1}^{n} (X_i - \hat{\mu}_X(n))^2 \leq c\right] = 0.95 \tag{6.5-1}$$

or, equivalently,

$$P\left[-a \leq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}} \leq a\right] \times P\left[b \leq \frac{1}{\sigma_X^2} \sum_{i=1}^{n} (X_i - \hat{\mu}_X(n))^2 \leq c\right] = 0.95. \tag{6.5-2}$$

Equation 6.5-2 follows from Equation 6.5-1 because of the independence of the events

$$E_1 \triangleq \left\{-a \leq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}} \leq a\right\} \text{ and } E_2 \triangleq \left\{b \leq \frac{1}{\sigma_X^2} \sum_{i=1}^{n} (X_i - \hat{\mu}_X(n))^2 \leq c\right\}.$$

We note that

$$Z \triangleq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}}$$

is the standard Normal RV $N(0,1)$ with distribution function $F_{SN}(z)$ while

$$W_n \triangleq \frac{1}{\sigma_X^2} \sum_{i=1}^{n} (X_i - \hat{\mu}_X(n))^2$$

is χ_{n-1}^2 with distribution function $F_{\chi^2}(x; n-1)$.

The next step is to associate a probability to each of the events E_1, E_2. As an example we could factor the joint δ-confidence as $\delta = \sqrt{\delta} \times \sqrt{\delta}$; this would give for $\delta = 0.95$

$$P\left[-a \leq \frac{\hat{\mu}_X(n) - \mu_X}{\sigma_X/\sqrt{n}} \leq a\right] = \sqrt{0.95} \simeq 0.975 \tag{6.5-3}$$

and

$$P\left[b \leq \frac{1}{\sigma_X^2}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \leq c\right] = \sqrt{0.95} \simeq 0.975. \tag{6.5-4}$$

From Equation (6.5-3) we recognize that $a = z_{0.9875}$, that is, $F_{SN}(z_{0.9875}) = 0.9875$, the 98.75 percentile of the standard Normal RV. From Equation 6.5-4, we determine–using the "equal-error" assignment rule to the tails of the Chi-square pdf–that $b = x_{0.0125}$ and $c = x_{0.9875}$, that is, the 1.25 and 98.75 percentiles of the cumulative Chi-square distribution $F_{\chi^2}(x; n-1)$. More generally, for any given δ-confidence interval and any given n, *we* can find numbers a, b, and c to satisfy the confidence constraints. Once this is done we can find in the μ, σ^2 parameter space the boundaries of the δ-confidence region for μ_X, σ_X^2. Event E_1 is the convex region *inside* the parabola described by $\sigma^2 = n\left(\mu - \hat{\mu}_X\right)^2/a^2$. Event E_2 is the region between the end points

$$\begin{aligned} \sigma_{\text{Max}}^2 &= \tfrac{1}{b}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \text{ (upper bound)}, \\ \sigma_{\text{Min}}^2 &= \tfrac{1}{c}\sum_{i=1}^{n}(X_i - \hat{\mu}_X(n))^2 \text{ (lower bound)}. \end{aligned} \tag{6.5-5}$$

The event $E_1 \cap E_2$ is then the shaded region shown in Figure 6.5-1.

In approximately 950 in a 1000 cases, the region shown in Figure 6.5-1 will cover the point μ_X, σ_X^2, that is, the true values of the unknown mean and variance.

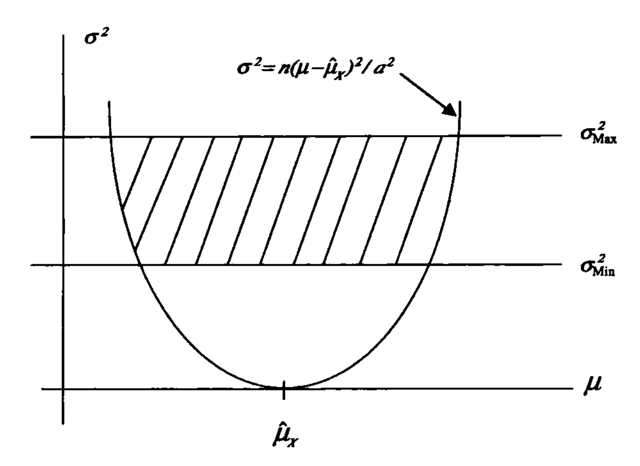


Figure 6.5-1 The confidence region for the combined estimation of μ and σ^2.

Example 6.5-1

(*confidence region for mean and variance*) We make 21 observations $\{X_i, i = 1, \ldots, 21\}$ on a Normal population $X : N(\mu_X, \sigma_X^2)$. A 90 percent confidence region is desired for the pair μ_X, σ_X^2.

To achieve a 90 percent confidence region, we assign (approximately) a 0.95 probability that the $N(0, 1)$ RV Z lies in the interval $(-a, a)$ and a 0.95 probability that the Chi-square Rv W with DOF of 20 lies in the interval (b, c). From Equation 6.5-3 we obtain $P[-z_{0.975} < Z \leq z_{0.975}] = 0.95$; hence, from the standard Normal distribution table, we find $F_{SN}(z_{0.975}) = 0.975$ or $z_{0.975} = 1.96$. From Equation 6.5-4 we obtain $P[b < W \leq c]$ $= 0.95$, from which we determine numbers $b = x_{0.025}, c = x_{0.975}$ using the "equal-error" assignment of Example 6.3-3. Thus, $F_{\chi^2}(x_{0.025}; 20) = 0.025$ and $F_{\chi^2}(x_{0.975}; 20) = 0.975$ so that $x_{0.025} = 9.59$ and $x_{0.975} = 34.2$. The numbers $x_{0.025}$ and$x_{0.975}$ are the 2.5 and 97.5 percentiles, respectively, of the χ^2 RV.

6.6 ESTIMATION OF NON-GAUSSIAN PARAMETERS FROM LARGE SAMPLES

Consider an RV X with mean μ and finite variance σ^2. We make n i.i.d. observations on $X\{X_i,\ i = 1, \ldots, n\}$ and deduce from the Central Limit Theorem that the *sample mean estimator*† (SME)

$$\hat{\mu}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is approximately Normal as $N(\mu, \sigma^2/n)$ for large n. If X is a continuous RV then the SME is approximately Normal in density, else it is approximately Normal in distribution. When the parameters to be estimated are associated with non-Gaussian distributions, it may still be possible to estimate them using Equation 6.6-1 as a starting point:

$$P\left[-a \leq \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \leq a\right] = \delta. \tag{6.6-1}$$

which can be rewritten as

$$P\Big[(-a\sigma/\sqrt{n}) + \hat{\mu} \leq \mu \leq (a\sigma/\sqrt{n}) + \hat{\mu}\Big] = \delta. \tag{6.6-2}$$

The reader will recognize that this is the expression for $100 \times \delta$ percent confidence interval for μ. When distributions are non-Gaussian, the mean and variance may be related parameters, that is, $\sigma = \sigma(\mu)$. How do we handle such cases? We illustrate with two examples from [6-2].

Example 6.6-1

(*confidence interval for λ in the exponential distribution*) Suppose we want to estimate λ in the exponential pdf $f_X(x) = \lambda e^{-\lambda x} u(x)$. For this law we find

†Recall we use the *mean-estimator function* and the *sample mean estimator* interchangeably.

$$\mu \triangleq E[X] = \int_0^\infty x\lambda e^{-\lambda x}dx = \lambda^{-1}$$

and

$$\sigma^2 \triangleq E[(X-\mu)^2] =) \int_0^\infty (X-\lambda^{-1})^2\lambda e^{-\lambda x}dx = \lambda^{-2}.$$

Inserting these results into Equation 6.6-2 and rearranging terms to expose λ yields

$$P\left[\frac{(-a/\sqrt{n})+1}{\hat{\mu}} \leq \lambda \leq \frac{(a/\sqrt{n})+1}{\hat{\mu}}\right] = \delta. \quad (6.6\text{-}3)$$

The number a is obtained from approximating $Z \triangleq (\hat{\mu}-\mu)\sqrt{n}/\sigma$ as a $N(0,1)$ random variable. This yields $a = z_{(1+\delta)/2}$ where $F_{SN}(z_{(1+\delta)/2}) = (1+\delta)/2$ Thus a $100 \times \delta$ percent confidence interval for λ has width

$$W_\delta = 2z_{(1+\delta)/2}/\hat{\mu}\sqrt{n} \quad (6.6\text{-}4)$$

Example 6.6-2

(*numerical evaluation of confidence interval for* $\boldsymbol{\lambda}$) It is desired to obtain a 95 percent confidence interval on the parameter λ of the exponential distribution from 64 i.i.d. observations on an exponential RV X. The estimate is $\hat{\mu}'_X = 3.5$. From Equation 6.6-1 we obtain $2\times\text{erf}(a) = 0.95$ or, equivalently, $F_{SN}(z_{(1+\delta)/2}) = (1+\delta)/2 = 0.975$. This gives $z_{0.975} = 1.96$. Then from Equations 6.6-3 and 6.6-4 we compute that the 95 percent confidence interval for λ is $\{0.22\ ,\ 0.36\}$ and has an approximate width of 0.14.

Example 6.6-3

(*confidence interval for p in the Bernoulli distribution*) Given a Bernoulli RV X, with probability $P[X=1]=p$, and $P[X=0]=q=1-p$ we want to estimate p at a $100 \times \delta$ percent level of confidence from n (sufficiently large) i.i.d. observations on X. For this distribution $\mu_X \triangleq E[X] = p$ and the MEF is $\hat{p} = (1/n)\sum_{i=1}^n X_i$. As demonstrated in earlier chapters $E[\hat{p}] = p$ and $\text{Var}[\hat{p}] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n^2}npq = pq/n$.

Hence the RV

$$Z \triangleq \frac{\hat{p}-p}{\sqrt{pq/n}} \quad (6.6\text{-}5)$$

for large n is Normal in *distribution* (not in density since X is a discrete RV) *as* $N(0,1)$. To obtain a $100 \times \delta$ confidence interval on p we write

$$P[-a \leq \frac{\hat{p}-p}{\sqrt{pq/n}} \leq a] = \delta \quad (6.6\text{-}6)$$

or, equivalently,

$$P[(p-\hat{p})^2 \leq a^2pq/n] = \delta.$$

As usual we find the constant a from $2\,\text{erf}(a) = \delta$, that is†, $a = z_{\frac{1+\delta}{2}}$ and find the end points of the confidence interval by solving for the roots of $(p-\hat{p})^2 - a^2pq/n = 0$(where $q = 1-p$). These are

†Recall that $2 \times \text{erf}(a) = 2 \times F_{SN}(a) - 1 = \delta$ so that $a = x_{(1+\delta)/2}$ i.e. the $(1+\delta)/2$ percentile of the standard Normal RV.

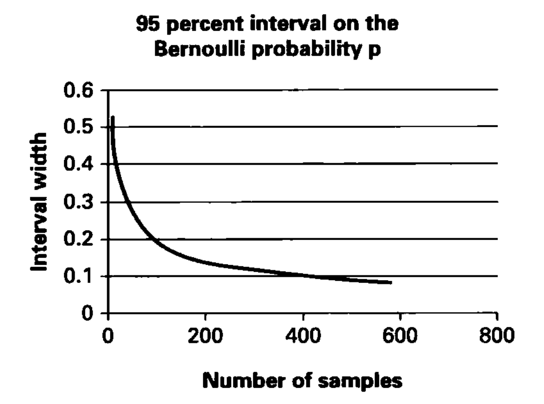


Figure 6.6-1 The width of the confidence interval decreases slowly with an increase in the number of samples in Example 6.6-3. Here we assumed that $4\hat{p}\hat{q} \approx 1$.

$$p_1 = \frac{2\hat{p} + (a^2/n)}{2(1 + (a^2/n))} - \frac{1}{2(1 + (a^2/n))}\sqrt{(a^2/n)[(a^2/n) + 4\hat{p}\hat{q}]}$$
$$p_2 = \frac{2\hat{p} + (a^2/n)}{2(1 + (a^2/n))} + \frac{1}{2(1 + (a^2/n))}\sqrt{(a^2/n)[(a^2/n) + 4\hat{p}\hat{q}]}$$

giving an interval width

$$W_{\gamma_a} = |p_2 - p_1| = \frac{1}{(1+(a^2/n))}\sqrt{(a^2/n)[(a^2/n) + 4\hat{p}\hat{q}]}. \tag{6.6-7}$$

The width of the interval decreases slowly with sample size Figure 6.6-1.

Example 6.6-4

(*how fair is the "fair" coin*) We wish to obtain information about the "fairness" of a coin. For this purpose the coin is tossed 100 times and 47 heads are observed. A 95 confidence interval on p, the probability of a head, is desired. Using the MEF we find that $\hat{p}' = 0.47$. We find a from $2 \times \text{erf}(a) = 0.95$ or $a = 1.96$ and from Equation 6.6-7, $W_\delta \simeq 0.192$. The interval is centered at 0.47 and extends from 0.37 to 0.57. The interval includes the "fair" coin value of $p = 0.5$ and we have no basis for believing that the coin is biased. If the number of i.i.d. observations increases to 1200, and we observe 564 heads, then $\hat{p}'$ still has value $\hat{p}' = 0.47$ but the 95 percent interval is $\{0.442, 0.492\}$ and does not include the "fair" value of 0.5. This strongly suggests that the coin has a slight bias in the direction of getting more tails.

6.7 MAXIMUM LIKELIHOOD ESTIMATORS

In the previous sections we furnished estimators for the mean, variance, and covariance of RVs. While these estimators enjoyed desirable properties, they seemed quite arbitrary in that they did not follow from any general principle. In this section, we discuss a somewhat general approach for finding estimators. This approach is called the *maximum likelihood*

(ML) principle and the estimators derived from it are called *maximum likelihood estimators* (MLEs). The main drawback to the MLE approach is that the underlying *form* of the pdf of the observed data must be known. The idea behind the MLE approach is illustrated in the following example.

Example 6.7-1

Consider a Bernoulli RV that has PMF $P_X(k) = p^k(1-p)^{1-k}$, where $P[X=1] = p$, and $P[X=0] = 1-p$. We would like to estimate the value of p with an estimator, say, $\hat{p}$, that is a function only of the observations on X. Suppose we make n observations on X and we call these observations $X_1, X_2, \ldots, X_n$. Then $Y = \sum_{i=1}^n X_i$ is the number of times that a *one* was observed in n tries. For example, the experiment might consist of tossing a coin n times and counting the number of times it came up heads, that is, $\{X=1\}$, when the probability of a head is p. Suppose this number is k_1. The *a priori* probability of observing k_1 heads is given by $P[Y=k_1;p] = \binom{n}{k_1} p^{k_1}(1-p)^{n-k_1}$. We explicitly show the dependence of the result on p because p is assumed unknown. We now ask what *value of p was most likely to have yielded this result*? Since the term on the right is a continuous function of p, we can obtain this result by a differentiation. Setting the derivative to zero yields

$$\frac{dP[Y=k_1;p]}{dp} = \binom{n}{k_1} p^{k_1-1}(1-p)^{n-k_1-1}[k_1(1-p) - p(n-k_1)] = 0.$$

Thus, there are three roots: $p=0$, $p=1$, and $p=k_1/n$. The first two roots yield a minimum while $p=k_1/n$ yields a maximum. Thus, our estimate for the most likely value of p in this case is k_1/n. Had we performed the experiment a second time and observed k_2 heads, our estimate for p would have been k_2/n. These estimates are realizations of the MLE for p:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{6.7-1}$$

In the previous example we used the fact that the distribution of $\sum_{i=1}^n X_i$ is binomial. Could we have obtained the same result without this knowledge? After all, for some distributions it might be quite a bit of work to compute the distribution of the sum of RVs. The answer is yes and the result is based on generation of the *likelihood function.*

Definition 6.7-1 The likelihood function† $L(\theta)$ of the random variables X_1, $X_2, \ldots, X_n$ is the joint pdf $f_{X_1X_2\ldots X_n}(x_1, x_2, \cdots, x_n; \theta)$ considered as a function of the unknown parameter θ. In particular if $X_1, X_2, \cdots, X_n$ are independent observations on a RV X with pdf $f_X(x;\theta)$, then the likelihood function for outcomes $X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i, \ldots, X_n = x_n$ becomes

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) \tag{6.7-2}$$

†Strictly speaking we should write $L(\theta; x_1, x_2, \ldots, x_n)$ or, as some books have, $L(\theta; X_1, X_2, \ldots, X_n)$. However, we dispense with this excessive notation.

since the $\{X_i\}$ are i.i.d. RVs with pdf $f_X(x;\theta)$. If, for a given outcome $X = (x_1, x_2, \cdots, x_n)$, $\theta^*(x_1, x_2, \cdots, x_n)$ is the value of θ that maximizes $L(\theta)$, then $\theta^*(x_1, x_2, \cdots, x_n)$ is the ML estimate of θ (a number) and $\hat{\theta} = \theta^*(X_1, X_2, \cdots, X_n)$ is the MLE (an RV) for θ. It is therefore, quite reasonable to define the likelihood function as the RV $L(\theta) \triangleq \prod_{i=1}^n f_X(X_i;\theta)$. Then, maximizing with respect to θ yields the MLE $\hat{\theta}(X_1, \cdots, X_n)$ directly. ■

Example 6.7-2

We consider finding the ML estimation of p in Example 6.7-1 using the likelihood function. If we make n i.i.d. observations $X_1, X_2, \cdots, X_n$ on a Bernoulli RV X, the likelihood function becomes $L(\theta) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} \times (1-p)^{n-\sum_{i=1}^n x_i}$. By setting $dL(\theta)/d\theta = 0$, we obtain three roots: $p = 0, p = 1$, and $p = \sum_{i=1}^n x_i/n$. The first two roots yield a minimum, while the last root yields a maximum. Thus, $p^*(\mathbf{x}) = \sum_{i=1}^n x_i/n$ and the MLE of p is $\hat{p} = p^*(X_1, X_2, \cdots, X_n) = \sum_{i=1}^n X_i/n$.

In many cases the differentiation is more conveniently done on the logarithm of the likelihood function. The *log-likelihood* function is $\log L(\theta)$ (usually the natural log is used) and has its maximum at the same value of θ as that of $L(\theta)$. Another point is that the MLE cannot always be found by differentiation, in which case we have to use other methods. Finally, *multiple-parameter ML estimation* can be done by solving simultaneous equations. We illustrate all three points in the next three examples, respectively.

Example 6.7-3

Assume $X{:}N(\mu, \sigma^2)$, where σ is known. Compute the MLE of the mean μ.

Solution The likelihood function for n realizations of X is

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right). \tag{6.7-3}$$

Since the log function is monotonic, the maximum of $L(\mu)$ is also that of $\log L(\mu)$. Hence

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$$

and set

$$\frac{\partial \log L(\mu)}{\partial \mu} = 0.$$

This yields

$$\sum_{i=1}^n (x_i - \mu) = 0.$$

Thus, the value of μ, say μ^*, that maximizes $L(\mu)$ is

$$\mu^* = \frac{1}{n}\sum_{i=1}^n x_i,$$

which implies that the MLE of μ should be

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6.7-4}$$

Thus, we see that in the Normal case, the MLE of μ can be computed by differentiation the log-likelihood function and that it turns out to be the sample mean.

Example 6.7-4

Assume X is uniform in $(0, \theta)$, that is,

$$f_X(x) = \begin{cases} \frac{1}{\theta}, 0 < x \le \theta, \\ 0, \ x > \theta, \end{cases}$$

and we wish to compute the MLE for θ. Let a particular realization of the n observations $X_1, \ldots, X_n$ be $\mathbf{x} = (x_1, \ldots, x_n)^T$ and let $x_m \triangleq \max(x_1, \ldots, x_n)$. The likelihood function is

$$L(\theta) = \begin{cases} \frac{1}{\theta^n}, x_m \le \theta, \\ 0, \quad \text{otherwise.} \end{cases}$$

Clearly to maximize L we must make the estimate θ' as small as possible. But θ' cannot be smaller than x_m. Hence θ' is x_m and the MLE is

$$\hat{\theta} = \max(X_1, \ldots, X_n). \tag{6.7-5}$$

The CDF of $\hat{\theta}$ for $n = 2$ is

$$F_{\hat{\theta}}(\alpha) = F_{X_1}(\alpha)F_{X_2}(\alpha) = F_X^2(\alpha). \tag{6.7-6}$$

We leave the computation of the CDF and pdf of $\hat{\theta}$ for arbitrary n as an exercise for the reader.

Example 6.7-5

Consider the Normal pdf

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad -\infty < x < \infty.$$

The log-likelihood function, for n realizations, is

$$\bar{L}(\mu, \sigma) \triangleq \log L = -\frac{n}{2}\log 2\pi - n\log\sigma$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{6.7-7}$$

Now set

$$\frac{\partial \bar{L}}{\partial \mu} = 0 \quad \frac{\partial \bar{L}}{\partial \sigma} = 0$$

and obtain the simultaneous equations

$$\sum_{i=1}^{n}(x_i - \mu) = 0 \tag{6.7-8}$$

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0. \tag{6.7-9}$$

From Equation 6.7-8 we infer that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6.7-10}$$

From Equation 6.7-9 we infer that, using the result from Equation 6.7-10,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2. \tag{6.7-11}$$

MLEs have a number of desirable properties including squared-error consistency and *invariance*. Invariance is that property that says that if $\hat{\theta}$ is the MLE for θ, then $h(\hat{\theta})$ is the MLE for $h(\theta)$. However, as seen in Example 6.7-5, (Equation 6.7-11) ML estimators cannot be counted on to be unbiased. We complete this section with an example that illustrates the invariance property.

Example 6.7-6

Consider n observations on a Normal RV. Assume that it is known that the mean is zero. The MLE of the variance is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$. The standard deviation σ is the square root of the variance. Hence the MLE of the standard deviation is the square root of the MLE for the variance, that is, $\hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{1/2}$.

6.8 ORDERING, MORE ON PERCENTILES, PARAMETRIC VERSUS NONPARAMETRIC STATISTICS

We make n i.i.d. observations on a generic RV X (recall that X is sometimes called a *population*) with CDF $F_X(x)$ to obtain the sample $X_1, X_2, \ldots, X_n$. The joint pdf of the sample is $f_X(x_1) \times \ldots \times f_X(x_n), -\infty < x_i < \infty,\ i = 1, \ldots, n$. Next we order the X_i, $i = 1, \ldots, n$, by size (signed magnitude) to obtain the *ordered sample* $Y_1, Y_2, \ldots, Y_n$ such that $-\infty < Y_1 < Y_2 < \cdots < Y_n < \infty$. When ordered, the sequence 3, $-2, -9$, 4 would become $-9, -2$, 3, 4. If a sequence $X_1 \ldots X_{20}$ was generated from n observations on $X : N(0,1)$, it would be very unlikely that $Y_1 > 0$ because this would require that the other 19 $Y_i, i = 2, \ldots, 20$, be greater than zero and therefore all the samples would be on the positive side of

the Normal curve. The probability of this event is $(1/2)^{20}$. Likewise it would be extremely unlikely that $Y_{20} < 0$ because this would require that the other 19 Y_i, $i = 1, \ldots, 19$ be less than zero. As shown in Section 5.3, the joint pdf of the ordered sample $Y_1, Y_2, \ldots, Y_n$ is $n! f_X(y_1) \times \cdots \times f_X(y_n), -\infty < y_1 < y_2 < \cdots < y_n < \infty$ and zero else. We distinguish between *ordering* and *ranking* in that ranking normally assigns a value to the ordered elements. For example, most people would order the pain of a broken bone higher than that of a sore throat due to a cold. But if a physician asked the patient to *rank* these pains on a scale of zero to ten, the pain associated with the broken bone might be ranked at eight or nine while the sore throat might be given a rank of three or four.

Consider next the idea of *percentiles*. We have already used this concept in numerous places in earlier discussions; here we elaborate. Assume that the IQ of a large segment of the population is distributed as $N(100, 100)$, that is, a mean of 100 and a standard deviation of 10. Obviously the Normal approximation is valid only over a limited range because no one has an IQ of 1000 or an IQ of –10. The IQ test itself is valid only over a limited range and may not give an accurate score for people that are extremely bright or severely cognitively handicapped. It is sometimes said that people in either group are "off the IQ scale." Still the IQ test is widely used as an indicator of problem-solving ability. Suppose that the result of an IQ test says that the child ranks in the 93rd percentile of the examinees and therefore qualifies for admission to selective schools. How do we locate the 93rd percentile in a population of n students?

Definition (percentile): Given an RV X with CDF $F_X(x)$, the u-percentile of X is the number x_u such that $F_X(x_u) = u$. If the function F_X is everywhere continuous with continuous derivative, then $x_u = F_X^{-1}(u)$, where F_X^{-1} is the inverse function associated with F_X, that is, $F_X^{-1}(F_X(x_u)) = x_u$. A CDF and its inverse function is shown in Figure 6.8-1. In keeping with common usage, we use x_u or $100 \times x_u$ interchangeably to mean x_u-percentile.

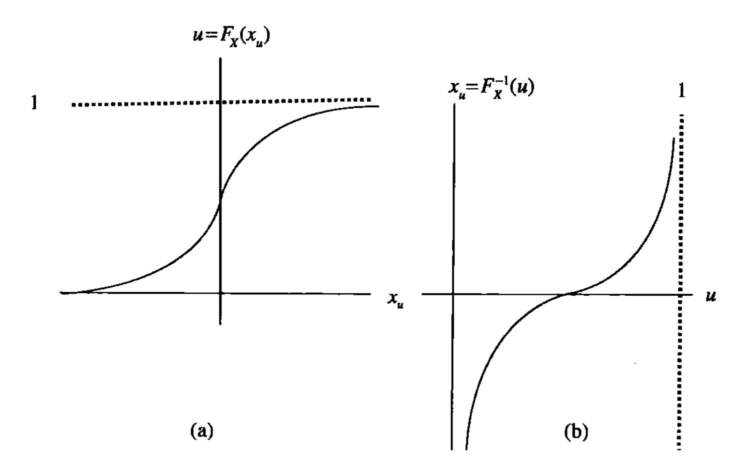


Figure 6.8-1 (a) u versus x_u; (b) The inverse function x_u versus u.

Observation. In the special case where $X{:}N(0,1)$ with CDF $F_{SN}(z)$, we use the symbol z_u (or $100 \times z_u$) to denote the u-percentile of X. If $X{:}N(\mu, \sigma^2)$ then the u-percentile of X, x_u, is related to z_u according to

$$x_u = \mu + z_u\sigma. \tag{6.8-1}$$

Example 6.8-1

(*relation between x_u and z_u*) We wish to show that $x_u = \mu + z_u\sigma$ if $X{:}N(\mu, \sigma^2)$. We proceed as follows:

We write

$$\begin{aligned} F_X(x_u) = u &= \left(2\pi\sigma^2\right)^{-1/2} \int_{-\infty}^{x_u} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right) dx \\ &= (2\pi)^{-1/2} \int_{-\infty}^{(x_u-\mu)/\sigma} \exp\left(-\frac{1}{2}z^2\right) dz \\ &\triangleq (2\pi)^{-1/2} \int_{-\infty}^{z_u} \exp\left(-\frac{1}{2}z^2\right) dz. \end{aligned}$$

The last line is $F_{SN}(z_u)$, the CDF of the standard Normal RV. Hence $x_u = \mu + z_u\sigma$. We can use this result in the previously mentioned IQ problem. From the data we have $F_X(x_u) = 0.93 = F_{SN}(z_u)$. We can find z_u from tables of the Normal CDF, or from tables of the error function $(\text{erf}(z_u) = F_{SN}(z_u) - 0.5)$ we get that $z_u \approx 1.48$. Then with $x_u = \mu + z_u\sigma$=100+1.48 (10), we get that a 93 percentile in the IQ is 115.

The Median of a Population Versus Its Mean

The *median* of the population X is the point $x_{0.5}$ such that $F_X(x_{0.5}) = 0.5$.† This is to be contrasted with the *mean* of X, written as μ_X, and defined as $\mu_X = \int_{-\infty}^{\infty} x f_X(x)dx$. The median and mean do not necessarily coincide. For example, in the case of the exponential law where $f_X(x) = \lambda e^{-\lambda x}u(x)$, we find that $\mu_X = 1/\lambda$ but $x_{0.5} = 0.69/\lambda$. To compute the mean of X we need $f_X(x)$, which is often not known. The mean may seem like a rather abstract parameter while the median is merely the point $x_{0.5}$ where $P[X \leq x_{0.5}]$. However, given n i.i.d. observations $X_1, X_2, \ldots, X_n$ on X, we *estimate* μ_X with the mean estimator function (MEF) $\hat{\mu}_X = n^{-1}\sum_{i=1}^{n} X_i$, which happens to be an unbiased and consistent estimator for the mean of many populations. Indeed it is the simple form of the MEF $\hat{\mu}_X$ and the fact that if σ_X^2 is finite that $\hat{\mu}_X \to \mu_X$ for large n (see the law of large numbers) that make the mean so useful in many applications. Realizations of the MEF are intuitively appealing as they give us a sense of the center of gravity of the data.

†When the event $\{X = x_{0.5}\}$ has zero probability, the events $\{X < x_{0.5}\}$ and $\{X > x_{0.5}\}$ are equally probable at 0.5. This gives rise to the often-heard statement that the median "is the point at which half the population is below and half above." But as the median is the 50th percentile, it includes the probability of the event $\{X = x_{0.5}\}$ and the statement should be modified to "the median is the point at which half the population is at or below." The median is a parameter that characterizes the whole population. The median of a random sample is only an estimate of the true median.

Example 6.8-2

(*median salary versus mean salary*) Consider a country where half the workers make \$10,000 per year or less and half make more. Then we can take \$10,000 as the *median annual income.* Now suppose that among those making \$10,000 or less per annum, the numerical-mean annual income is \$8000 while for those making more than \$10,000 per annum, the *numerical-mean annual income* is \$100,000. The numerical mean income for the country as a whole in this case is \$54,000. In your judgment, which of these figures describes the economy of the country better? Which of these figures would you use to put the country in a good (bad) light?

Example 6.8-3

(*median and mean are not the same for the binomial*) We make the somewhat trivial observation that in the binomial case the mean and median do not coincide. For example, with $n = 5$, the mean is 2.5 but the median, such as it is, is 2. However, when n is large, the median and mean approach each other and the median can be estimated by the mean. Indeed stated without proof, the difference between the mean and median is proportional to $(p(1-p))^n$, which becomes arbitrarily small for $n \to \infty$.

Parametric versus Nonparametric Statistics

The situation where we know or assume a functional form for a density, distribution, or probability mass function and use this information in computing probabilities, estimating parameters, and making decisions is called the *parametric statistics.* Typically, in the *parametric case*, we might assume a form for the population density, for example, the Normal, and wish to estimate some unknown parameter of the distribution, for example, the mean μ_X. In Chapter 7 we make extensive use of parametric statistics in hypothesis testing. Much of parametric statistics is based on the Central Limit Theorem, which states that the distribution of the sum of a large number of i.i.d. observations tends to the Normal CDF.

The estimation of the properties and parameters of a population without any assumptions on the form or knowledge of the population distribution is known as *distribution-free* or *nonparametric statistics.* Statistics based only on observations without assuming underlying distributions are sometimes said to be *robust* in the sense that the theorems and conclusions drawn from the observations do not change with the form of the underlying distributions. Whereas the mean and standard deviation are useful in characterizing the center and dispersion of a population in the parametric case, the *median* and *range* play a comparable role in the nonparametric case. To estimate the median from $X_1, X_2, \ldots, X_n$, we order them by magnitude as $Y_1 < Y_2 < \ldots < Y_n$ and estimate $x_{0.5}$ with the *sample median* estimator

$$\hat{Y}_{0.5} = \begin{cases} Y_{k+1} & \text{if } n \text{ is odd, that is, } n = 2k+1, \\ 0.5(Y_k + Y_{k+1}) & \text{if } n \text{ is even, that is, } n = 2k. \end{cases} \tag{6.8-2}$$

The sample median is not an unbiased estimator for $x_{0.5}$ but becomes nearly so when n is large. The dispersion in the nonparametric case is measured from the 50 percent range, that is, $\Delta x_{0.50} \triangleq x_{0.75} - x_{0.25}$, or the 90 percent range, that is, $\Delta x_{0.90} \triangleq x_{0.95} - x_{0.05}$ or some other appropriate range. These have to be estimated from the observations.

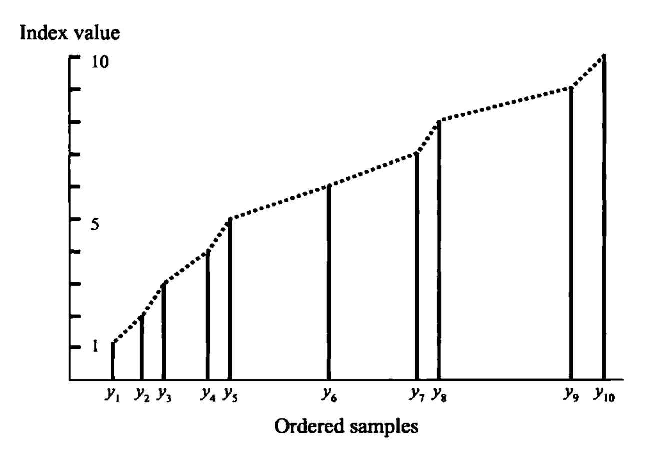


Figure 6.8-2 Estimated percentile range from ten ordered samples showing linear interpolation between the samples. To get the estimated percentile, take the index value and multiply by 100/11. Thus, to a first approximation, the 90th percentile is estimated from y_{10} while the 9th percentile is estimated from y_1. An approximate 50 percent range is covered by $y_8 - y_2$.

Example 6.8-4

(*interpolation to get percentile points*) Using the symbol $\alpha \sim \beta$ to mean α estimates β, we have $Y_3 \sim x_{0.273}, Y_4 \sim x_{0.364}$ and using linear interpolation

$$Y_4 + \frac{(Y_4 - Y_3)(0.3 - 4/11)}{1/11} \sim x_{0.3}.$$

Interpolation between samples is shown in Figure 6.8-2.

Confidence Interval on the Percentile

We discuss next a fundamental result connecting order statistics with percentiles. Once again the model is that of collecting a sample of n i.i.d. observations $X_1, X_2, \ldots, X_n$ on a RV X with CDF $F_X(x)$. We recall the notation $P[X_i \leq x_u] \triangleq u$. Next we order the samples by signed magnitude to get $Y_1 < Y_2 < \cdots < Y_n$. To remind the reader: if a set of realizations of the $X_i, i = 1, \ldots, 5$, are $x_1 = 7$, $x_2 = -2$, $x_3 = 7.2$, $x_4 = 1$, $x_5 = 3$ then the associated realizations on the $Y_i, i = 1, \ldots, 5$, are $y_1 = -2, y_2 = 1, y_3 = 3, y_4 = 7, y_5 = 7.2$. From the subscripts on $\{Y_i\}$ we can make an obvious but remarkable statement on the $\{X_i\}$, namely that the event $\{Y_k < x_u\}$ implies that there are *at least* k of the $\{X_i\}$ that are less than x_u; there may be more but certainly not less. Then, because the $\{X_i\}$ are i.i.d., we can use the binomial probability formula to compute $P[Y_k < x_u]$ as

$$\begin{aligned} P[Y_k < x_u] &= P[\text{at least } k \text{ of the } \{X_i\} \text{ are less than } x_u] \\ &= \textstyle\sum_{i=k}^{n} \binom{n}{i} u^i (1-u)^{n-i}. \end{aligned} \tag{6.8-3}$$

Next consider the event $\{Y_{k+r} > x_u\}$. Since Y_{k+r} is the $(k+r)$th element in the ordering of the $\{X_i\}$, there are *at least* $n-(k+r)+1$ of the $\{X_i\}$ that are greater than x_u. Equivalently there can be *no more* than $k+r-1$ of the $\{X_i\}$ less than x_u. Then

$$P[Y_{k+r} > x_u] = P[\text{no more than } k+r-1 \text{ of the } \{X_i\} \text{ are less than } x_u] \\ = \sum_{i=0}^{k+r-1} \binom{n}{i} u^i (1-u)^{n-i}. \tag{6.8-4}$$

The intersection of the events $\{Y_{k+r} > x_u\}$ and$\{Y_k < x_u\}$ is the event $\{Y_k < x_u < Y_{k+r}\}$. Its probability is

$$P[Y_k < x_u < Y_{k+r}] = \sum_{i=k}^{k+r-1} \binom{n}{i} u^i (1-u)^{n-i} \tag{6.8-5}$$

and is independent of $f_X(x)$. The result given in Equation 6.8-5 is one of the major results of nonparametric statistics and has important applications as we illustrate below.

Example 6.8-5

(*sample size needed to cover the median at 95 percent confidence*) We seek the end points Y_1, Y_n of a random interval $[Y_1, Y_n]$ so that the event $\{Y_1 < x_{0.5} < Y_n\}$ occurs with probability 0.95. Here $Y_1 \triangleq \min(X_1, X_2, \ldots X_n), Y_n \triangleq \max(X_1, X_2, \ldots X_n)$. In effect, how large should n be?

The answer is furnished by computing

$$P[Y_1 < x_{0.5} < Y_n] = \sum_{i=1}^{n-1} \binom{n}{i} (1/2)^n \approx 0.95$$

and find that for $n=5$, $P[Y_1 < x_{0.5} < Y_5] \approx 0.94$. The probability that the random interval $[Y_1, Y_n]$ covers the 50 percent percentile point is shown in Figure 6.8-3 for various values of n.

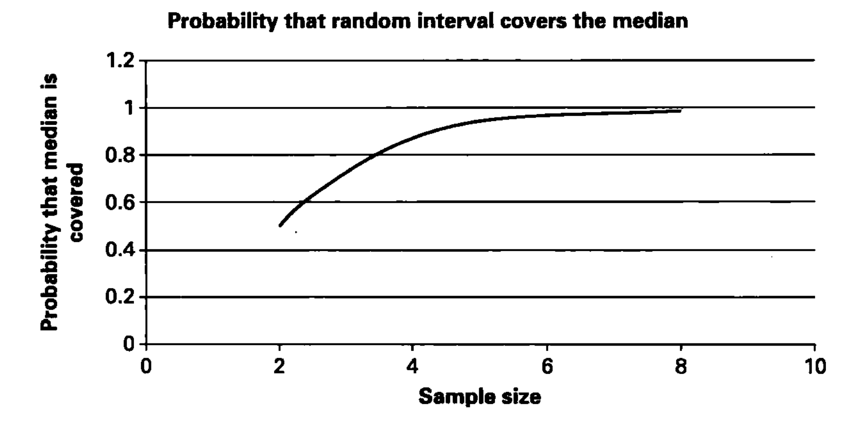


Figure 6.8-3 Probability that the event $\{Y_1 < x_{0.5} < Y_n\}$ covers the median for various values of n.

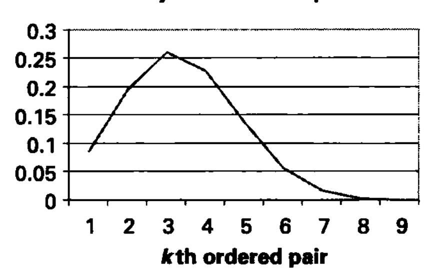


Figure 6.8-4 Among the pairwise intervals $[Y_k, Y_{k+1}]$, the interval $[Y_3, Y_4]$ is most likely to cover $x_{0.33}$.

Example 6.8-6

(*between which pair of ordered samples does $x_{0.33}$ lie?*) We have a set of ordered samples $\{Y_1, Y_2, \ldots, Y_n\}$ and wish to find the pair $\{Y_i, Y_{i+1}, i = 1, \ldots, n-1\}$ that maximizes the probability of covering the 33.33rd percentile point. The 33.33rd percentile point is defined by $u = 1/3 = F_X(x_{0.33})$. For specificity we assume $n = 10$. From Equation 6.8-5 we compute

$$P[Y_k < x_{0.33} < Y_{k+1}] = \frac{10!}{k!(10-k)!}(1/3)^k(2/3)^{10-k}, k = 1, \ldots, 9$$

and plot the result in Figure 6.8-4. Clearly the interval $[Y_3, Y_4]$ is most likely to cover $x_{0.33}$. The probability of the event $\{Y_3 < x_{0.33} < Y_4\}$ is 0.26.

Confidence Interval for the Median When n Is Large

If n is large enough so that the Normal approximation to the binomial is valid in distribution, we can use

$$P[\alpha \leq S_n \leq \beta] \approx \frac{1}{\sqrt{2\pi}} \int_{\alpha_n}^{\beta_n} \exp\left[-\frac{1}{2}y^2\right] dy, \tag{6.8-6a}$$

where

$$P[\alpha \leq S_n \leq \beta] = \sum_{i=\alpha}^{\beta} \binom{n}{i} p^i (1-p)^{n-i},$$

$$\alpha_n \triangleq \frac{\alpha - np - 0.5}{\sqrt{np(1-p)}}, \text{ and} \tag{6.8-6b}$$

$$\beta_n \triangleq \frac{\beta - np + 0.5}{\sqrt{np(1-p)}}.$$

To apply these results to the problem at hand, we write

$$P[Y_r < x_{0.5} < Y_{n-r+1}] = \sum\nolimits_{i=r}^{n-r} \binom{n}{i} (1/2)^n, \tag{6.8-7}$$

where we used that, by definition of the median, $u = F_X(x_{0.5}) = 1/2$. The choice of subscripts will ensure that the confidence interval will begin at the rth place counting from the bottom, that is, from one, and end at the place reached by counting r observations back from the top. For example if the 95 percent confidence calculation for $n = 10$ yields $r = 3$, the confidence interval begins at the third observation and ends at the eighth observation, both points reached by counting three places from bottom and top, respectively, that is, 1, 2, 3 (Y_3) and 10, 9, 8 (Y_8), and the result would appear as $P[Y_3 < x_{0.5} < Y_8] = 0.95$.

In the binomial sum in Equation 6.8-7 we note that its mean is $n/2$ and its standard deviation is $\sqrt{n}/2$. Hence the Normal approximation to the binomial sum in Equation 6.8-7 for a 95 percent confidence interval is

$$\sum_{i=r}^{n-r} \binom{n}{i}(1/2)^n \approx \frac{1}{\sqrt{2\pi}} \int_{\alpha_n}^{\beta_n} \exp[-\frac{1}{2}x^2]dx = 0.95,$$

which, from the tables of the standard Normal distribution (or the error function), yields $\alpha_n = -1.96, \beta_n = 1.96$. Then it follows from Equation 6.8-6b that

$$1.96 = \frac{n - r - n/2 + 0.5}{\sqrt{n}/2}$$

$$-1.96 = \frac{r - n/2 - 0.5}{\sqrt{n}/2},$$

which yields $r = (n/2) - 1.96\sqrt{n}/2 + 0.5$. If r is not an integer replace r by $\lfloor r \rfloor$, where the latter is the largest integer less than or equal to r.

Example 6.8-7

(*95 percent confidence interval for the median for n=20*) We make 20 observations on an RV X and label these $\{X_i, i = 1, \ldots, 20\}$. We order them by signed magnitude so that $Y_1 < Y_2 < \cdots < Y_n$. We use $r = (n/2) - 1.96\sqrt{n}/2 + 0.5$ to obtain $r = 6.12$ and $\lfloor r \rfloor = 6$. Then $P[Y_6 < x_{0.5} < Y_{15}] \geq 0.95$.

6.9 ESTIMATION OF VECTOR MEANS AND COVARIANCE MATRICES†

Let $\mathbf{X}_1 \triangleq (X_i, \ldots, X_p)^T$ be a p-component random vector with pdf $f_X(x)$. Let $X_i, \ldots, X_n$ be n observations on X, that is, the $X_i, i = 1, \ldots, n$ are drawn from $f_X(x)$. Then $X_i, i = 1, \ldots n$ are i.i.d. random vectors with pdf $f_X(x_i)$. We show below how to estimates

$$\text{(i) } \boldsymbol{\mu}_\mathbf{X} \triangleq E[\mathbf{X}] = (\mu_1, \ldots, \mu_p)^T,$$

†This section and the next one can be omitted on a first reading.

where

$$\mu_j \triangleq E[X_j] \qquad j = 1, \ldots, p$$

and

$$\text{(ii) } \mathbf{K_{XX}} \triangleq E[(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})^T].$$

The vector and matrix parameter $\boldsymbol{\mu}_\mathbf{X}$ and $\mathbf{K_{XX}}$ are useful in many signal processing applications. They also figure prominently in characterizing the multi-dimensional Normal distribution [6-5]. The covariance matrix $\mathbf{K_{XX}}$ is most often a full-rank, positive-definite, real-symmetric matrix. The properties of such matrices are well-known [6-6] and can be exploited in their estimation.

Estimation of μ

Consider the p-vector estimator $\hat{\boldsymbol{\Theta}}$ given by

$$\hat{\boldsymbol{\Theta}} \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i. \tag{6.9-1}$$

We shall show that $\hat{\boldsymbol{\Theta}}$ is unbiased and consistent for μ. We arrange the observations as in Table 6.9-1.

In Table 6.9-1 X_{ij} is jth component of the random vector X_i. The components of the vector $Y_i, j = 1, \ldots, p$ are n i.i.d. observations on the jth component of the random vector X. From the scalar case we already know that

$$\hat{\boldsymbol{\Theta}}_j \triangleq \frac{1}{n} \sum_{i=1}^{n} X_{ij} \triangleq \hat{\mu}_j \quad j = 1, \cdots, p \tag{6.9-2}$$

Table 6.9-1 Observed Data

	$\mathbf{X}_1 \ldots \mathbf{X}_i \ldots \mathbf{X}_n$
$\mathbf{Y}_1$	X_{11} X_{i1} X_{n1}
$\vdots$	. $\vdots$.
$\mathbf{Y}_j$	. X_{ij} . p rows
$\vdots$	. $\vdots$.
$\mathbf{Y}_p$	X_{1p} X_{np} (n columns)

The components of $\mathbf{Y}_j$ are all that is necessary for estimating the jth component, μ_j, of the vector μ.

is unbiased and consistent for $\mu_j \triangleq E[X_{ij}]\ i = 1, \cdots, n$. It follows therefore that the vector estimator $\hat{\mathbf{\Theta}} \triangleq (\hat{\Theta}_1, \cdots, \hat{\Theta}_p)^T$ is unbiased and consistent for μ. The vector Y_j contains all the information for estimating μ_j. Thus, $E[Y_j] = \mu_j i$, where $i \triangleq (1, 1, \cdots, 1, 1)^T$.

When X is normal, $\hat{\mathbf{\Theta}}$ is normal. Even when X is not normal, $\hat{\mathbf{\Theta}}$ tends th the normal for large n by the central limit theorem (Theorem 4.7-1).

Estimation of the covariance *K*

If the mean μ is known, then the estimator

$$\hat{\mathbf{\Theta}} \triangleq \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T \tag{6.9-3}$$

is unbiased for $\mathbf{K}$. However, since the mean is generally estimated from the *sample mean* $\hat{\mu}$, it turns out that the estimator

$$\hat{\mathbf{\Theta}} \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \hat{\boldsymbol{\mu}})(\mathbf{X}_i - \hat{\boldsymbol{\mu}})^T \tag{6.9-4}$$

is unbiased for $\mathbf{K_{XX}}$. To prove this result requires some effort. First observe that the diagonal elements of $\hat{\mathbf{\Theta}}$ are of the form

$$S_{jj} \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \hat{\mu}_j)^2, \tag{6.9-5}$$

which we already know from the univariate case are unbiased for $\sigma_j^2 \triangleq E[(X_j - \mu_j)^2]$. Next consider the sequence $(l \neq m)$

$$X_{1l} + X_{1m}, X_{2l} + X_{2m}, \cdots, X_{nl} + X_{nm}, \tag{6.9-6}$$

which are n i.i.d. observations $Z_{lm}^{(i)}$, on a univariate RV $Z_{lm} \triangleq X_l + X_m$ with mean $\mu_l + \mu_m$ and variance

$$\begin{aligned}\text{Var}[Z_{lm}] &= E[(X_l - \mu_l) + (X_m - \mu_m)]^2 \\ &= \sigma_l^2 + \sigma_m^2 + 2K_{lm}\end{aligned} \tag{6.9-7}$$

where $K_{lm} \triangleq E[(X_l - \mu_l)(X_m - \mu_m)]$ is the lmth element of $\mathbf{K_{XX}}$. Finally, consider

$$\hat{\mathbf{\Theta}}_{lm} \triangleq \frac{1}{n-1}\sum_{l=1}^{n}[Z_{lm}^i - (\hat{\mu}_l + \hat{\mu}_m)]^2, \tag{6.9-8}$$

which, by Equation 6.8-15, is unbiased for $\sigma_l^2 + \sigma_m^2 + 2K_{lm}$. If we expand Equation 6.9-8 and use the fact that $Z_{lm} \triangleq X_l + X_m$, we obtain

$$\begin{aligned}\hat{\Theta}_{lm} &\triangleq \frac{1}{n-1}\sum_{i=1}^{n}[(X_{il}-\hat{\mu}_l)+(X_{im}-\hat{\mu}_m)]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_{il}-\hat{\mu}_l)^2 + \frac{1}{n-1}\sum_{i=1}^{n}(X_{im}-\hat{\mu}_m)^2 \\ &+ \frac{2}{n-1}\sum_{i=1}^{n}(X_{il}-\hat{\mu}_l)(X_{im}-\hat{\mu}_m). \end{aligned} \tag{6.9-9}$$

In Equation 6.9-9, the first term is unbiased for σ_l^2, the second is unbiased for σ_m^2, and the sum of all three is unbiased by Equation 6.9-8 for $\sigma_l^2 + \sigma_m^2 + 2K_{lm}$. We therefore conclude that

$$S_{lm} \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(X_{il}-\hat{\mu}_l)(X_{im}-\hat{\mu}_m) \tag{6.9-10}$$

is unbiased for $K_{lm}(= K_{ml})$. Hence every term of $\hat{\Theta}$ in Equation 6.9-4 is unbiased for every corresponding term in $\mathbf{K_{XX}}$. In this sense $\hat{\Theta} \triangleq \hat{\mathbf{K}}_{\mathbf{XX}}$ is unbiased for $\mathbf{K_{XX}}$.

By resorting again to the univariate case and assuming that all moment up to the fourth order exist, we can show consistency for every term in the estimator for $\mathbf{K_{XX}}$, that is Equation 6.9-4. Hence without specifying the distribution, Equations 6.9-1 and 6.9-4 are unbiased and consistent estimators for $\boldsymbol{\mu}_\mathbf{X}$ and $\mathbf{K_{XX}}$ respectively.

When $\mathbf{X}$ is normal, $\hat{\mathbf{K}}_{\mathbf{XX}}$ obeys a structurally complex probability law called the Wishart distribution (see 6-6, p. 126). More generally, when the *form* of the pdf of $\mathbf{X}$ is known, one can use the maximimum likelihood method of estimatiing such parameters as $\boldsymbol{\sigma}_\mathbf{X}^2, \boldsymbol{\mu}_\mathbf{X}$ and $\mathbf{K_{XX}}$. Maximum likelihood estimators have several, but not all, desirable properties as estimators. The next example shows that the MLE for the mean is not a minimum-square estimator.

Example 6.9-1

([6-5], p. 21.) Consider the sample mean estimator from Equation 6.8-3, that is,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We recognize that this estimator is the MLE for the mean μ' Now we ask: what constant a in the scalar estimator $\hat{\Theta} \triangleq a\hat{\mu}$ will generate the MMSE estimator of μ? Recall the X_i $i = 1, \cdots, n$ are i.i.d. RV with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$.

Solution we are seeking the value of a such that

$$E[a\hat{\mu} - \mu]^2 \tag{6.9-11}$$

is a minimum. Clearly $\hat{\mu}$ is unbiased for μ, and it seems hard to believe that there may exist. an $\hat{\Theta}$ with $a \neq 1$ that—though yielding a biased estimator—gives a lower MSE than $\hat{\Theta} = \mu$.

For *any* estimator $\hat{\Theta}$, the mean square error in estimating μ is

$$E[(\hat{\Theta} - \mu)^2] = E[\{(\hat{\Theta} - E[\hat{\Theta}]) + (E[\hat{\Theta}] - \mu)\}^2]$$
$$= \text{Var}[\hat{\Theta}] + (E[\hat{\Theta}] - \mu)^2. \tag{6.9-12}$$

If $\hat{\Theta}$ is unbiased then the last term, which is the square of the bias (Definition 6.8-2), is zero. For the case at hand, $\hat{\Theta} = a\hat{\mu}$; thus

$$E[(\hat{\Theta} - \mu)^2] = a^2\text{Var}[\hat{\mu}] + (a\mu - \mu)^2$$
$$= \frac{a^2\sigma^2}{n} + (a-1)^2\mu^2. \tag{6.9-13}$$

To find the MMSE estimator, we differentiate Equation 6.9-13 with respect to a and set to zero. This yields the optimum value of $a = a_0$, that is,

$$a_0 = \frac{\mu^2}{(\sigma^2/n) + \mu^2} = \frac{n}{(\sigma^2/n) + n}, \tag{6.9-14}$$

6.10 LINEAR ESTIMATION OF VECTOR PARAMETERS†

Many measurement problems in the real world are described by the following model:

$$y(t) = \int_T h(t, \tau)\theta(\tau)d\tau + n(t), \tag{6.10-1}$$

where $y(t)$ is the *observation* or *measurement*, T is the integration set, $\theta(\tau)$ is the unknown *parameter* function, $h(t, \tau)$ is a function that is characteristic of the system and links the parameter function to the measurement but is itself independent of $\theta(\tau)$, and $n(t)$ is the inevitable error in the measurement due to noise. For computational purposes Equation 6.10-1 must be reduced to its discrete form

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N}, \tag{6.10-2}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations with components Y_i, $i = 1, \ldots, n$. $\mathbf{H}$ is a known $n \times k$ matrix $(n > k), \boldsymbol{\theta}$ is an unknown $k \times 1$ parameter vector, and $\mathbf{N}$ is an $n \times 1$ random vector whose unknown components $N_i, i = 1, \ldots, n$ are the errors or noise associated with the ith observation Y_i. We shall assume without loss of generality that $E[\mathbf{N}] = 0$.‡

†This section can be omitted on a first reading.

‡The symbol 0 here stands for the zero vector, that is, the vector whose components are all zero.

Equation 6.10-2 is known as the *linear model.* We now ask the following question: How do we extract a "good" estimate of $\boldsymbol{\theta}$ from the observed values of $\mathbf{Y}$ if we restrict our estimator $\hat{\Theta}$ to be a linear function of $\mathbf{Y}$? By a linear function we mean

$$\hat{\boldsymbol{\Theta}} = \mathbf{BY}, \tag{6.10-3}$$

where $\mathbf{B}$, which *does not* depend on $\mathbf{Y}$, is to be determined. The problem posed here is of practical significance. It is one of the most fundamental problems in parameter estimation theory and covered in great detail in numerous books, for example, Kendall and Stuart [6-8] and Lewis and Odell [6-9]. It also is an immediate application of the probability theory of random vectors and is useful for understanding various topics in subsequent chapters.

Before computing the matrix $\mathbf{B}$ in Equation 6.10-3, we must first furnish some results from matrix calculus.

Derivative of a scalar function of a vector. Let $q(\mathbf{x})$ be a scalar function of the vector $\mathbf{x} = (x_1, \ldots, x_n)^T$. Then

$$\frac{dq(\mathbf{x})}{d\mathbf{x}} \triangleq \left(\frac{\partial q}{\partial x_1}, \ldots, \frac{\partial q}{\partial x_n}\right)^T. \tag{6.10-4}$$

Thus, the derivative of $q(\mathbf{x})$ with respect to $\mathbf{x}$ is a *column vector* whose ith component is the partial derivative of $q(\mathbf{x})$ with respect to x_i.

Derivative of quadratic forms. Let $\mathbf{A}$ be a real-symmetric $n \times n$ matrix and let $\mathbf{x}$ be an arbitrary n-vector. Then the derivative of the quadratic form

$$q(\mathbf{x}) \triangleq \mathbf{x}^T\mathbf{A}\mathbf{x}$$

with respect to $\mathbf{x}$ is

$$\frac{dq(\mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x}. \tag{6.10-5}$$

The proof of Equation 6.10-5 is obtained by writing

$$\begin{aligned} q(\mathbf{x}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j \\ &= \sum_{i=1}^{n} x_i^2 a_{ii} + \sum_{i \neq j}^{n}\sum^{n} a_{ij} x_i x_j. \end{aligned}$$

Hence

$$\begin{aligned} \frac{\partial q(\mathbf{x})}{\partial x_k} &= 2x_k a_{kk} + 2\sum_{i \neq k} a_{ki} x_i \\ &= 2\sum_{i=1}^{n} a_{ki} x_i \end{aligned}$$

or

$$\frac{dq(\mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x}. \tag{6.10-6}$$

Derivative of scalar products. Let $\mathbf{a}$ and $\mathbf{x}$ be two n-vectors. Then with $y = \mathbf{a}^T\mathbf{x}$, we obtain

$$\frac{dy}{d\mathbf{x}} = \mathbf{a}. \tag{6.10-7}$$

Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{A}$ be two n-vectors and an $n \times n$ matrix, respectively. Then with $q \triangleq \mathbf{y}^T\mathbf{A}\mathbf{x}$,

$$\frac{\partial q}{\partial \mathbf{x}} = \mathbf{A}^T\mathbf{y}. \tag{6.10-8}$$

We return now to Equation 6.10-2:

$$\mathbf{Y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{N}$$

and assume that (recall $E[\mathbf{N}] = \mathbf{0}$)

$$\mathbf{K} \triangleq E[\mathbf{N}\mathbf{N}^T] = \sigma^2\mathbf{I} \tag{6.10-9}$$

where $\mathbf{I}$ is the identity matrix. Equation 6.10-9 is equivalent to stating that the measurement errors N_i, that is, $i = 1, \ldots, n$ are uncorrelated, and their variances are the same and equal to σ^2. This situation is sometimes called *white noise.*

A reasonable choice for estimating $\boldsymbol{\theta}$ is to find a $\hat{\Theta}$ that *minimizes* the sum squares S defined by

$$S \triangleq (\mathbf{Y} - \mathbf{H}\hat{\mathbf{\Theta}})^T(\mathbf{Y} - \mathbf{H}\hat{\mathbf{\Theta}}) \triangleq ||\mathbf{Y} - \mathbf{H}\hat{\mathbf{\Theta}}||^2. \tag{6.10-10}$$

Note that by finding $\hat{\mathbf{\Theta}}$ that best fits the measurement $\mathbf{Y}$ in the sense of minimizing $||\mathbf{Y} - \mathbf{H}\hat{\Theta}||^2$, we are realizing what is commonly called a *least-squares* fit to the data. For this reason, finding $\hat{\mathbf{\Theta}}$ that minimizes S in Equation 6.10-10 is called the least-squares (LS) method. It is a form of the MMSE estimator. To find the minimum of S with respect to $\hat{\mathbf{\Theta}}$, write

$$S = \mathbf{Y}^T\mathbf{Y} + \hat{\mathbf{\Theta}}^T\mathbf{H}^T\mathbf{H}\hat{\mathbf{\Theta}} - \hat{\mathbf{\Theta}}^T\mathbf{H}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{H}\hat{\mathbf{\Theta}}$$

and compute (use Equation 6.10-4 on the LHS and Equations 6.10-5 and 6.10-8 on the RHS)

$$\frac{\partial S}{\partial \hat{\mathbf{\Theta}}} = 0 = 2[\mathbf{H}^T\mathbf{H}]\hat{\mathbf{\Theta}} - 2\mathbf{H}^T\mathbf{Y},$$

whence (assuming $\mathbf{H}^T\mathbf{H}$ has an inverse)

$$\hat{\mathbf{\Theta}}_{LS} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Y}. \tag{6.10-11}$$

Comparing our result with Equation 6.10-3 we see that the $\mathbf{B}$ in Equation 6.10-3 that furnishes the least-squares solution is given by $\mathbf{B}_0 \triangleq (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$. Equation 6.10-11 is the LS estimator of $\boldsymbol{\theta}$ based on the measurement $\mathbf{Y}$.

The astute reader will have noticed that we never involved the fact that $\mathbf{K} = \sigma^2\mathbf{I}$. Indeed, in arriving at Equation 6.10-11 we essentially treated $\mathbf{Y}$ as deterministic and merely obtained $\hat{\mathbf{\Theta}}_{LS}$ as the *generalized inverse* (see Lewis and Odell [6-9, p. 6]) of the system of equations $\mathbf{Y} = \mathbf{H}\boldsymbol{\theta}$. As it stands, the estimator $\hat{\mathbf{\Theta}}_{LS}$ given in Equation 6.10-11 has no claim to being optimum. However, when the covariance of the noise $\mathbf{N}$ is as in Equation 6.10-9, then $\hat{\mathbf{\Theta}}_{LS}$ does indeed have optimal properties in an important sense. We leave it to the reader to show that $\hat{\mathbf{\Theta}}_{LS}$ is unbiased and is a minimum variance estimator.

Example 6.10-1

We are given the following data

$$6.2 = 3\theta + n_1,$$

$$7.8 = 4\theta + n_2,$$

$$2.2 = \theta + n_3.$$

Find the LS estimate of θ.

Solution The data can be put in the form

$$\mathbf{y} = \mathbf{H}\boldsymbol{\theta} + \mathbf{n},$$

where $\mathbf{y} = (6.2, 7.8, 2.2)^T$ is a realization of $\mathbf{Y}$, $\mathbf{H}$ is a column vector described by $(3, 4, 1)^T$ and $\mathbf{n} = (n_1, n_2, n_3)^T$ is a realization of $\mathbf{N}$. Hence $\mathbf{H}^T\mathbf{H} = \Sigma H_i^2 = 26$ and $\mathbf{H}^T\mathbf{y} = \Sigma_{i=1}^3 H_i y_i = 52$. Thus,

$$\hat{\theta}_{LS} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} = \frac{\sum_{i=1}^{3} H_i y_i}{\sum_{i=1}^{3} H_i^2} = \frac{52}{26} = 2.$$

Example 6.10-2

([6-8, p. 77.]) Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ be a two-component parameter vector to be estimated, and let $\mathbf{H}$ be a $n \times 2$ matrix of coefficients partitioned into column vectors as $\mathbf{H} = (\mathbf{H}_1\mathbf{H}_2)$, where $\mathbf{H}_i$, $i = 1, 2$ is an n-vector. Then with the n-vector $\mathbf{Y}$ representing the observation data, the linear model assumes the form

$$\mathbf{Y} = (\mathbf{H}_1\mathbf{H}_2)\boldsymbol{\theta} + \mathbf{N}$$

and the LS estimator of $\boldsymbol{\theta}$ is

$$\hat{\mathbf{\Theta}}_{LS} = \begin{bmatrix} \mathbf{H}_1^T\mathbf{H}_1 & \mathbf{H}_1^T\mathbf{H}_2 \\ \mathbf{H}_2^T\mathbf{H}_1 & \mathbf{H}_2^T\mathbf{H}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{H}_1^T\mathbf{Y} \\ \mathbf{H}_2^T\mathbf{Y} \end{bmatrix}.$$

SUMMARY

In the branch of statistics known as parameter estimation we apply the tools of probability to observational data to estimate parameters associated with probability functions. We began the chapter by stressing the importance of independent, identically distributed (i.i.d.) observations on a random variable of interest. We then described how these observations can be organized to estimate parameters such as the mean and variance, with emphasis on the Normal distribution. The problem of making "hard" (i.e., categorical) statements about parameters when the number of observations is finite was resolved using the notion of *confidence intervals*. Thus, we were able to say that based on the observations, the true mean, or variance, or both had to lie in a computed interval with a near 100 percent confidence. We studied the properties of the standard mean-estimating function and found that it was unbiased and consistent.

We found that the t-distribution, describing the probabilistic behavior of the T random variable, was of central importance in constructing a confidence interval for the mean of a Normal random variable when the variance is unknown.

In estimating the variance of a Normal random variable, we found that the Chi-square distribution was useful in constructing a near 100 percent confidence interval for the variance. We briefly discussed a method of estimating the standard deviation of a Normal random variable from ordered observations.

We demonstrated that confidence intervals could also be developed for parameters of distributions other than the Normal. This was demonstrated with examples from the exponential and Bernoulli distributions.

A method of estimating parameters based on the idea of which parameter was most likely to have produced the observational data was discussed. This method, *called maximum likelihood estimation* (MLE), is very powerful but does not always yield unbiased or minimum mean-square error estimators.

Toward the end of the chapter we introduced nonparametric methods for parameter estimation. These methods, also called *distribution-free estimation*, do not assume a specific distribution for generating the observational data. In this sense they are said to be *robust.* We found that a number of important results in the nonparametric case could be obtained using ordered data and the binomial distribution.

Finally, we extended the earlier discussions on parameter estimation to the vector case. In particular, we showed how the elements of vector means and covariance matrices could be estimated from observational data. A brief discussion of estimating vector parameters from linear operation on measurement data completed the chapter.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

6.1 If 36 of 100 persons interviewed are familiar with tax incentives for installing certain energy saving devices, construct a 95% confidence interval for the corresponding true proportion. What is the margin of error?

6.2 We have three i.i.d. observations on $X{:}N(0,1)$. Call these X_i, $i = 1,2,3$. Compute $f_{X_1X_2X_3}(x_1,x_2,x_3)$ and compare with $f_{X_1+X_2+X_3}(y)$.

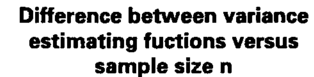
6.3 In a village in a developing country, 361 villagers are exposed to the Ebola Gay hemorrhaging-fever virus. Of the 361 exposed villagers, 189 die of the virus infection. Compute a 95 percent confidence interval on the probability of dying from the Ebola virus once you have been exposed to it. What is the margin of error?

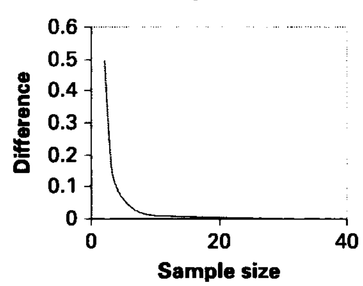
6.4 Show that the roots of the polynomial $(p-\hat{p})^2 - (9/n)p(1-p) = 0$ that appeared in Example 6.1-6 are indeed as given in Equation 6.1-1.

6.5 Referring again to Example 6.1-6, compute $|p_1 - p_2|$ as $\hat{p}$ varies from zero to one. Do this for different values of n, for example, $n = 0, 20, 30, 50$.

6.6 Describe how you would test for the fairness of a coin with a 95 percent confidence interval on the probability that the coin will come up heads.

6.7 Consider the variance estimating functions in Equations 6.3-3 and 6.3-4. Show that for values of $n \geq 20$, the difference between them becomes extremely small. Reproduce the curve shown below.





6.8 Compute $P[|\hat{\mu}_X(n) - \mu_X| \leq 0.1]$ as a function of n when $X : N(1,1)$.

6.9 Plot the width of a 95 percent confidence interval on the mean of a Normal random variable whose variance is unity versus the number of samples n.

6.10 Show that the MGF of the gamma pdf

$$f(x;\alpha,\beta) = \left(\alpha!\beta^{(\alpha+1)}\right)^{-1} x^{\alpha}\exp(-x/\beta), x > 0;\ \alpha > -1, \beta > 0$$

is $M(t) = (1-\beta t)^{-(\alpha+1)}$.

6.11 We make n i.i.d. observations X_i $i = 1,\ldots,n$ on $X : N(\mu,\sigma^2)$ and construct $Y_i = \dfrac{X_i - \mu}{\sigma}$. Use the result of Problem 6.10 to show that the pdf of $W_n \triangleq \sum_{i=1}^{n} Y_i^2$ is χ^2, with n degrees of freedom that is,

$$f_W(W;n) = ([(n/2)-1]!2^{n/2})^{-1}w^{(n/2)-1}\exp(-(1/2)w), w > 0.$$

6.12 We make n i.i.d. observations X_i on $X{:}N(\mu,\sigma^2)$ and construct $\hat{\mu} = n^{-1}\sum_{i=1}^{n} X_i$ and $\hat{\sigma}^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$. Show that $\hat{\mu}$ and $\hat{\sigma}^2$ are independent. (Hint: It helps to use moment generating functions; if all else fails consult Appendix F).

6.13 Let $X{:}N(\mu, \sigma^2)$ and $W_n{:}\chi_n^2$ be two independent RVs. With $Y \triangleq \dfrac{X-\mu}{\sigma}$:

a) Show that the joint dencity of Y and W_n is given by:

$$f_{YW_n}(y,w) = \frac{1}{\sqrt{2\pi}}\exp(-0.5y^2) \times \frac{W^{(n-2)/2}\exp(-0.5w)}{[(n-2)/2]!2^{n/2}}, -\infty < y < \infty, w > 0;$$

b) Let $T \triangleq \dfrac{Y}{\sqrt{W_n|n}}$ show that $f_T(t;n) = \dfrac{[(n-1)/2]!}{\sqrt{n\pi}[(n-2)/2]!} \times \dfrac{1}{[1+t^2/n]^{(n+1)/2}}$, $-\infty < t < \infty$. This the "Student's" t-pdf. (Hint use a proper two variable-to-two variable transformation.)

6.14 Let $(X_1, X_2, \ldots, X_n)$ be a random sample of a uniform random variable X over $(0, a)$, where a is unknown. Show that $A = \max(X_1, X_2, \ldots, X_n)$ is a consistent estimator of the parameter a.

6.15 Use Matlab™, Excel™, or some other scientific computing program to create a 95 percent confidence interval for the mean of a Normal random variable $X : N(0,1)$. Use 50 observations per single interval computation and repeat the experiment 50 times. For each experiment record the length of the interval and whether it includes the mean, which in this case is zero. Repeat for 100 observations per interval computation.

6.16 Show that the sample variance in Equation 6.2-3 is unbiased.

6.17 Show that the sample variance in Equation 6.2-3 is consistent.

6.18 Suppose that we want to estimate the true proportion of defectives in a very large shipment of adobe bricks, and that we want to be at least 95% confident that the error is at most 0.04. How large a sample will we need if we know that the true proportion does not exceed 0.12?

6.19 Consider a box that contains a mix of red and blue balls whose exact composition of red and blue balls is not known. If we draw n balls from the box with replacement and obtain k red balls, what is the maximum likelihood estimate of p, the probability of drawing a red ball?
Find a 95 percent confidence interval for the variance σ_X^2 of the distribution.

6.20 Show that the number a in Equation 6.6-6 is $a = z_{(1+\delta)/2}, -a = z_{(1-\delta)/2}$, that is, a is the $(1+\delta)/2$ percentile of the $Z{:}N(0,1)$.

6.21 An optical firm purchases glass to be ground into lenses. It knows from past experience that the variance of the refractive index of this kind of glass is 1.26×10^{-4}. Suppose that the refractive indices of 20 pieces of glass (randomly selected from a large shipment purchased by the optical firm) have a variance of 1.20×10^{-4}. Construct a 95% confidence interval for σ, the standard deviation of the population sampled.

6.22 Let X_1, X_2, X_3 be three observations on $X{:}N(\mu_X, \sigma_X^2)$. Let $V_i \triangleq \frac{X_i - \hat{\mu}_X(n)}{\sigma_X}$ for $i = 1, 2, 3$. Show that $\sum_{i=1}^3 V_i^2$ is Chi-square with two degrees of freedom.

6.23 A random sample of size $n = 100$ is taken from a population with $\sigma = 5.1$. Given that the sample mean is $\bar{x} = 21.6$, construct a 95% confidence interval for the population mean μ. Since the sample size $n = 100$ is large, normal approximation is assumed.

6.24 A method has been developed to estimate the size of the fish population by performing a capture/recapture experiment. Let N be the actual population to be estimated. r animals are first captured and tagged. The r animals are then released and allowed to mix into the general population. Later, n animals are captured (or recaptured) and the number of tagged animals k is counted. Determine the maximum likelihood estimate of N.

6.25 Show that the covariance estimating function of Equation 6.3-1 is unbiased and consistent.

6.26 Find the mean and variance of the random variable X driven by the geometric probability mass function $P_X(n) = (1-a)a^n u(n)$. Compute a 95 percent confidence interval on the mean of X.

6.27 In Example 6.5-3 the claim is made that $P\left[-a \leq \frac{\hat{p}-p}{\sqrt{pq/n}} \leq a\right] = \delta$ is identical with $P[(p-\hat{p})^2 \leq a^2pq/n] = \delta$. Justify this claim.

6.28 Compute the maximum likelihood estimate for the parameter λ in the Poisson pmf.

6.29 Let X be uniformly distributed in $(-1,1)$ and let $Y = X^2$. Find the best linear estimator for Y in terms of X^2. Compare its performance to the best estimator.

6.30 Compute the MLE for the parameter $p \triangleq P[success]$ in the binomial PMF.

6.31 Compute the MLE for the parameters a, b $(a < b)$ in $f_X(x) = (b-a)^{-1}(u(x-a)-u(x-b))$.

6.32 [6-2] Consider the linear model $\mathbf{Y} = \mathbf{Ia} + \mathbf{bx} + \mathbf{V}$, where

$$\begin{aligned}
\mathbf{Y} &\triangleq (Y_1, \ldots, Y_n)^T \\
\mathbf{V} &\triangleq (V_1, \ldots, V_n)^T \\
\mathbf{I} &\triangleq n \times n \text{ identity matrix} \\
\mathbf{x} &\triangleq (x_1, \ldots, x_n)^T \\
\mathbf{a} &\triangleq (a, \ldots, a)^T
\end{aligned}$$

and a, b are constants to be determined. Assume that the $\{V_i,\, i = 1, \ldots, n\}$ are n i.i.d. Normal random variables as $N(0, \sigma^2)$, the $x_i, i = 1, \ldots, n$ are constant for each $i = 1, \ldots, n$, but may vary as i varies. They are called *control variables.*

(i) Show that $\{Y_i, i = 1, \ldots, n\}$ are $N(a + bx_i, \sigma^2)$;

(ii) Write the likelihood function and argue that it is maximized when $\sum_{i=1}^n (Y_i - (a + bx_i))^2$ is minimized;

(iii) Show that the MLE of a is $\hat{a}_{ML} = \hat{\mu}_Y - \hat{b}_{ML}\bar{x}$ and the MLE of b is $\hat{b}_{ML} = \dfrac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$, where

$$\begin{aligned}
\hat{\mu}_Y &\triangleq (1/n)\textstyle\sum_{i=1}^n Y_i \\
\bar{x} &\triangleq (1/n)\textstyle\sum_{i=1}^n x_i.
\end{aligned}$$

6.33 The mean height of students in a college follows a Normal distribution with mean 173.3 cm and standard deviation of 6.4 cm. Determine the 95th percentile of the height random variable.

6.34 Compute the median for the geometrically distributed RV.

6.35 Compute the mean and median for the Chi-square random variable.

6.36 Show that an estimate for the 30th percentile, $x_{0.3}$, is given by the interpolation formula $Y_4 + \frac{(Y_4-Y_3)(0.3-4/11)}{1/11} \sim x_{0.3}$, where the $\{Y_i,\ i = 1, \ldots, 10\}$ are the ordered random variables formed from the set of unordered i.i.d random variables $\{X_i,\ i = 1, \ldots, 10\}$.

6.37 How large a sample do we need to cover the 50th percentile with probability 0.99? Hint: Use the formula $P[Y_1 < x_{0.5} < Y_n] = \sum_{i=1}^{n-1} \binom{n}{i} (1/2)^n \approx 0.99$, where $Y_1 \triangleq \min(X_1, X_2, \ldots X_n), Y_n \triangleq \max(X_1, X_2, \ldots X_n)$.

***6.38** Show that the joint pdf of the ordered random variables Y_1, Y_n, where $Y_1 \triangleq \min(X_1, X_2, \ldots X_n), Y_n \triangleq \max(X_1, X_2, \ldots X_n)$, is given by

$$f_{Y_1 Y_n}(y_1, y_n) = n(n-1)\left(F_X(y_n) - F_X(y_1)\right)^{n-2} f_X(y_1) f_X(y_n), -\infty < y_1 < y_n < \infty$$

Hint: Consider the joint pdf of all the $Y_1, Y_2, \cdots, Y_n$ and integrate out all but the first and last.

***6.39** Let $\{Y_i,\ i = 1, \ldots, n\}$ be a set of ordered random variables. Define the *range* R of the set as $R \triangleq Y_n - Y_1$. Now consider six observations on $X\{X_i, i = 1, \ldots, 6\}$ from the pdf $f_X(x) = u(x) - u(x-1)$, where $u(x)$ is the unit-step function. Show that $f_R(r) = 30r^4(1-r), 0 < r < 1$. Hint: Use the result $f_{Y_1 Y_n}(y_1, y_n) = n(n-1)\left(F_X(y_n) - F_X(y_1)\right)^{n-2} f_X(y_1) f_X(y_n), -\infty < y_1 < y_n < \infty$, and define two random variables $R \triangleq Y_n - Y_1, S \triangleq Y_1$ and find $f_{RS}(r, s)$. Then integrate out with respect to S.

REFERENCES

6-1. A. M. Mood and F. A. Graybill, *Introduction to the Theory of Statistics*, 2nd Edition. New York: McGraw-Hill, 1963.

6-2. A. Papoulis, *Probability & Statistics*. Englewood Cliffs, NJ. Prentice Hall, 1990.

6-3. H. Stark and J. G. Brankov, "Estimating the Standard Deviation from Extreme Gaussian Values," *IEEE Signal Processing Letters*, Vol. 11, No. 3, 2004, pp. 320–322.

6-4. W. Luo, "A Comment on 'Estimating the Standard Deviation from Extreme Gaussian Values'," *IEEE Signal Processing Letters*, Vol. 12, No. 2, 2005, p. 109.

6-5. K. S. Miller, *Multidimensional Gaussian Distributions*. New York: John Wiley, 1964.

6-6. J. N. Franklin, *Matrix Theory*. Upper Saddle River, NJ. Prentice Hall, 1968.

6-7. K. Fukunaga, *Introduction to Statistical Pattern Recognition*. 2nd edition. New York: Academic, 1990.

6-8. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics,* Vol.2, 3rd Edition. London, England: Charles Griffin and Co., 1951.

6-9. T. O. Lewis and P. L. Odell, *Estimation in Linear Models,* Upper Saddle River, NJ. Prentice-Hall, 1971.

ADDITIONAL READING

A. L. Garcia, *Probability, Statistics and Random Processes for Electrical Engineering,* 3rd Edition, Upper Saddle River, NJ. Prentice Hall, 2008.

S. M. Kay, *Fundamentals of Signal Processing,* Upper Saddle River, NJ. Prentice Hall, 1993.

G. Straud, *Linear Algebra and its Applications,* 3rd Edition. New York: Saunders, 1900.

(Online) Y. Cohen and J. Y Cohen, Chapter 9 in *Statistics and data with R: An Applied Approach through Examples.* New York: Wiley Online Library, 2008.

J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London, Series A*, Vol. 231, 1933, pp. 289–337.

T. A. Severini, "On the Relationship Between Bayesian Interval Estimates," *Journal of the Royal Statistical Society, Series B,* Vol. 53, No. 3, 1991, pp. 611–618.

J. E. Bernardo and A. Smith, *Bayesian Theory.* New York: Wiley, p. 259.

(Online) Tutor Vista.com: Probability Calculator

(Online) Statistics Help@ Talk Stats Forum

7 Statistics: Part 2 Hypothesis Testing

Hypothesis testing is an important topic in the broader area of statistical decision theory. Statistical decision theory also includes other activities such as prediction, regression, game theory, statistical modeling, and many elements of signal processing. However, the ideas underlying hypothesis testing often serve as the basis for these other, typically more advanced, areas.[†]

Hypotheses take the following form: We make a hypothesis that a parameter has a certain value, or lies in a certain range, or that a certain event has taken place. The so-called alternative hypothesis[‡] is that the parameter has a different value, or lies in a different range, or that an event has not taken place. Then, based on real data, we either accept (reject) the hypothesis or accept (reject) the alternative. Parameter estimation and hypothesis testing are clearly related. For example, the decision to accept the hypothesis that the mean of one population is equal to the known mean of another population is essentially equivalent to estimating the mean of the unknown population and deciding that it is close enough to the given mean to deem them equal.

In the real world we often are forced to make decisions when we don't have all the facts, or when our knowledge comes from observations that are inherently probabilistic. We all (probably) know heavy smokers who live well into their eighties and beyond. Likewise, we know of nonsmokers that die of lung cancer in their fifties. Does this mean that smoking is unrelated to lung cancer? In days of old, the chiefs of tobacco companies said *yes* while

[†]There are several textbook references for this material, for example [7-1] to [7-4].

[‡]The alternative hypothesis is often called, simply, the *alternative*. Thus, one encounters "we test the hypothesis... versus the alternative...."

cancer epidemiologist said *no*. In view of all the evidence accumulated since then, no reasonable person would now argue that smoking does not increase the likelihood of dying from lung cancer. Nevertheless, unlike what happens when a person falls off a 20-story building onto concrete, death by lung cancer or other smoking-related disease does not always follow heavy smoking. The relationship between smoking and lung cancer remains essentially probabilistic. In the following sections we discuss strategies for decision making in a probabilistic environment.

7.1 BAYESIAN DECISION THEORY

In the absence of divine guidance, the Bayesian approach to making decisions in a random (stochastic) environment is, arguably, the most rational procedure devised by humans. Unfortunately, to use Bayes in its original form requires information we may not have with any accuracy or may be impossible to get. We illustrate the application of Bayesian theory and its concomitant weakness in the following example.

Example 7.1-1

(*deciding whether to operate or not*) Assume that you are a surgeon and that your patient's x-ray shows a nodule in his left lung. The patient is 40 years old, has no history of smoking, and is otherwise in good health. Let us simplify the problem and assume that there are only two possible states: (1) The nodule is an early onset cancer that without treatment will spread and kill the patient and (2) the nodule is benign and doesn't pose a health risk. We shall abbreviate the former by the symbol ζ_1 and the latter by ζ_2. The reader will recognize that the outcome space (read *sample space*) Ω has only the two points, that is, $\Omega = \{\zeta_1, \zeta_2\}$, but—in more complex situations—could in fact have many more. The surgeon's job is to make that *decision* (and take subsequent *action*) that is *best* for the patient. The trouble is that without an operation the surgeon doesn't know the state of nature, that is, whether ζ_1 or ζ_2 is the case. There are two terminal actions: operate (a_1) or don't operate (a_2).

It is not always clear as to what "best" means. However, it seems quite reasonable, other things being equal, that "best" in this case is that decision/action that will minimize the number of years that the patient will lose from a normal lifetime. There are four situations to consider:

(1) The surgeon decides not to operate and the nodule is benign;
(2) The surgeon decides not to operate and the nodule is a cancer;
(3) The surgeon operates and the nodule is benign;
(4) The surgeon operates and the nodule is a cancer.

Prior data exist that lung nodules discovered in nonsmoking, early middle-age males are benign 70 percent of the time. Thus, the probability that a nodule is cancerous for this group is only 0.3. The surgeon is also aware of the data in Table 7.1-1.

The terms $\{l(a_i, \zeta_j), i = 1,2; j = 1,2\}$ are called *loss functions* and $l(a_i, \zeta_j)$ *is the loss* associated with taking action a_i when the state of nature is ζ_j. The reader might ask why $l(a_1, \zeta_1) = l(a_1, \zeta_2) = 5$ and not zero. Surgeons know that operations are risky

Table 7.1-1

If the decision is	And the state of nature is	Then the number of years subtracted from a normal life span is $l(a,\zeta)$
Don't operate (action a_2)	Benign lesion (ζ_2)	$l(a_2,\zeta_2)=0$
Don't operate (action a_2)	Cancer (ζ_1)	$l(a_2,\zeta_1)=35$
Operate (action a_1)	Benign lesion (ζ_2)	$l(a_1,\zeta_2)=5$
Operate (action a_1)	Cancer (ζ_1)	$l(a_1,\zeta_1)=5$

procedures and that even healthy patients can suffer from post-operative infections such as from MRSA† and gram-negative bacteria‡. Unless absolutely necessary, most surgeons will avoid major invasive surgery in preference to non-invasive procedures. Thus, due to infections and other complications any surgery carries a risk and, counting the people who die from surgical complications, we assign an average loss of five years.

Next, we introduce the idea of a *decision function d*. The decision function d is a function of observable data so we write $d(X_1, X_2, \ldots X_n)$, where the $\{X_i, i = 1, \ldots, n\}$ are n i.i.d. observations on a random variable (RV) X. The decision function $d(X_1, X_2, \ldots X_n)$ helps to guide the surgeon with respect to what action, that is, a_1 or a_2, to take. In our example we limit ourselves to a single observation that we denote X, specifically the ratio of the square of the length of the boundary of the nodule to the enclosed area. This is a measure of the *irregularity* of the edges of the nodule: The more irregular the edges, the more likely that the nodule is a cancerous lesion (Figure 7.1-1). Thus, we expect that most of the time the RV X for the cancerous lesion (ζ_1) will yield larger realizations than those yielded by X for the benign case (ζ_2). A realization of X in this case is the *datum*.

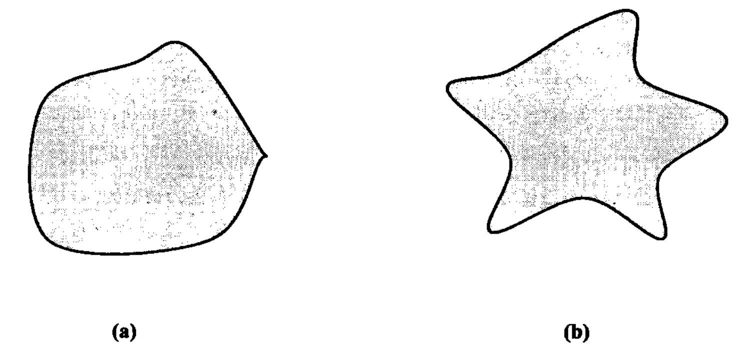


Figure 7.1-1 (a) A benign lesion tends to have regular edges; (b) a cancerous lesion tends to have irregular edges.

†Methicillin-resistant *Staphylococcus aureus*.

‡These bugs prevail in hospitals and cause infections that are difficult to treat.

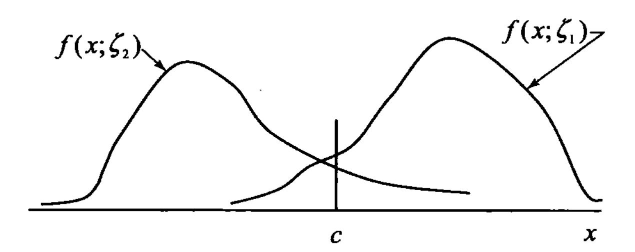


Figure 7.1-2 There is a value of c (to be determined) that will minimize the expected risk. A datum point in the region $\Gamma_2 \triangleq (-\infty, c]$ is more likely to be associated with a benign condition and will lead to action a_2 (don't operate), while a datum point in $\Gamma_1 \triangleq [c, \infty)$ is more likely to be associated with a cancer and will lead to action a_1 (operate).

Let $f(x;\zeta_1)$ and $f(x;\zeta_2)$ denote the pdf's of X under conditions ζ_1 and ζ_2, respectively (see Figure 7.1-2). In this example we assume, for simplicity and ease of visualization, that these pdf's are unimodal and are continuous. Further, as shown in Figure 7.1-2, we assume that there exists a constant c such that if the datum falls to the right of c it will be taken as evidence that the opacity is a cancer. Likewise, if the datum falls to the left of c it will be taken as evidence that cancer is not present. If the evidence suggests a cancer then action a_1 follows; else, action a_2 follows. Since this is a probabilistic environment errors will be made. Thus,

$$P[a_1|\zeta_2] = \int_c^{\infty} f(x;\zeta_2)dx \tag{7.1-1}$$

is the error probability that the evidence suggests there is a cancer that requires an operation when in fact there is no cancer. Likewise

$$P[a_2|\zeta_1] = \int_{-\infty}^{c} f(x;\zeta_1)dx \tag{7.1-2}$$

is the error probability that the evidence suggests there is no cancer and therefore the action is not to operate while in fact there is a cancer.

The conditional expectation of the loss when the state of nature is ζ and the decision rule is d is called the *risk* $R(d;\zeta)$. Thus,

$$R(d;\zeta_1) = l(a_1;\zeta_1)P[a_1|\zeta_1] + l(a_2;\zeta_1)P[a_2|\zeta_1]$$

and (7.1-3)

$$R(d;\zeta_2) = l(a_1;\zeta_2)P[a_1|\zeta_2] + l(a_2;\zeta_2)P[a_2|\zeta_2].$$

Finally, the *expected risk*, labeled $B(d)$ defined as†

$$B(d) = R(d;\zeta_1)P[\zeta = \zeta_1] + R(d;\zeta_2)P[\zeta = \zeta_2], \tag{7.1-4}$$

†The symbol B is used in honor of the mathematician/philosopher Tomas Bayes (1702–1761).

is the quantity to be minimized. A decision function d^* that minimizes $B(d)$ is a *Bayes strategy*.

Thus,

$$B(d^*) = \min_d \left\{ R(d;\zeta_1)P[\zeta = \zeta_1] + R(d;\zeta_2)P[\zeta = \zeta_2] \right\}. \tag{7.1-5}$$

The probabilities

$$P_1 \triangleq P[\zeta = \zeta_1], P_2 \triangleq P[\zeta = \zeta_2]$$

are called the *a priori* probabilities of the state of nature. In terms of the symbols introduced above, we can write $B(d)$ as

$$B(d) = P_1 \times l(a_2,\zeta_1) + P_2 \times l(a_2,\zeta_2)$$
$$+ \int_c^\infty \left\{ P_2 f(x;\zeta_2)\left[l(a_1,\zeta_2) - l(a_2,\zeta_2)\right] - P_1 f(x;\zeta_1)\left[l(a_2,\zeta_1) - l(a_1,\zeta_1)\right] \right\} dx, \tag{7.1-6}$$

where we choose c to minimize $B(d)$. If the integral in the expression for $B(d)$ is positive, it will add to $B(d)$, but if the integral is negative, it will reduce $B(d)$. Indeed if we choose c, say $c = c^*$, so that (c^*, ∞) leaves out all the points where the integral is positive but includes the points where the integral is negative, then we have minimized $B(d)$. Outcomes (read *events*) that make the integral negative are described by

$$\frac{f(X;\zeta_1)}{f(X;\zeta_2)} > \frac{\left[l(a_1,\zeta_2) - l(a_2,\zeta_2)\right] P_2}{\left[l(a_2,\zeta_1) - l(a_1,\zeta_1)\right] P_1} \triangleq k_b, \tag{7.1-7}$$

which is the *Bayes decision rule*. It says that for all outcomes† (c^*, ∞) take action a_1 (operate). Likewise for all outcomes $(-\infty, c^*)$, that is,

$$\frac{f(X;\zeta_1)}{f(X;\zeta_2)} < k_b,$$

take action a_2 (don't operate). The constant c is the point that satisfies

$$\frac{f(c^*;\zeta_1)}{f(c^*;\zeta_2)} = k_b.$$

The prior probabilities in this example would be computed from aggregate information on thousands of patients who sought help for similar symptoms. The nodule observed in a 40-year, nonsmoking male is more likely to be *benign* than *cancerous*; for example, it might be a harmless opacity, some residual scar tissue, or even the intersection of blood vessels giving the appearance of a lesion. For simplicity we shall assume that we know these probabilities as $P_1 = 0.7, P_2 = 0.3$. Then specializing Equation 7.1-6 for this case yields

$$B(d) = 10.5 + \int_c^\infty (3.5f(x;\zeta_2) - 9f(x;\zeta_1))\, dx,$$

†Recall that under the mapping of the (real) RV X, events are intervals on the real line.

which implicitly yields the constant c^* from $3.5f(c^*;\zeta_2) - 9f(c^*;\zeta_1) = 0$. Then the Bayes decision rule is

$$f(X;\zeta_1)/f(X;\zeta_2) > 0.39 \rightarrow \text{operate}$$
$$f(X;\zeta_1)/f(X;\zeta_2) < 0.39 \rightarrow \text{don't operate.}$$

In Example 7.1-1 only a single RV was used in making the decision. In many problems, however, a decision will be based on observing many i.i.d. RVs. In that case the Bayes decision rule takes the form

$$\frac{f(X_1;\zeta_1)\cdots f(X_n;\zeta_1)}{f(X_1;\zeta_2)\cdots f(X_n;\zeta_2)} > \frac{[l(a_1,\zeta_2)-l(a_2,\zeta_2)]\,P_2}{[l(a_2,\zeta_1)-l(a_1,\zeta_1)]\,P_1} \triangleq k_b, \text{ accept } \zeta_1 \text{ as state of nature}$$

$$\frac{f(X_1;\zeta_1)\cdots f(X_n;\zeta_1)}{f(X_1;\zeta_2)\cdots f(X_n;\zeta_2)} < \frac{[l(a_1,\zeta_2)-l(a_2,\zeta_2)]\,P_2}{[l(a_2,\zeta_1)-l(a_1,\zeta_1)]\,P_1} \triangleq k_b, \text{ reject } \zeta_1 \text{ as state of nature.}$$

(7.1-8)

The reader will recognize that the numerator and denominator in Equation 7.1-8 are the likelihood functions $L(\zeta_j) = \prod_{i=1}^{n} f_X(X_i;\zeta_j)$, $j = 1,2$ discussed in Chapter 6. Therefore Equation 7.1-8, being a ratio of two likelihood functions (*the likelihood ratio*) that is being compared to a constant, is quite appropriately called a *likelihood ratio test* (LRT). The constant k_b in Equation 7.1-8 is called the *Bayes threshold.*

Every Bayes strategy leads to an LRT but not every LRT is the result of a Bayes strategy. The Bayes strategy seeks to *minimize the average risk* but other LR-type tests may seek to abide by different criteria, for example, maximizing the LR subject to a given probability of error. One problem with implementing the Bayes strategy is that the *a priori* probabilities P_1 and P_2 are often not known. Another problem is that it may be difficult to assign a reasonable "loss" to a particular action. For example, say that you are preparing a large omelet and need to break a dozen eggs. You are thinking of using a Bayes strategy to minimize the loss, that is, the amount of work you have to do. Your choices are to use one bowl or two bowls and the random element here is whether an egg is good or bad. Suppose that, on average, for every 100 good eggs there is one bad egg. If you use only one bowl and a bad egg is added to the others before you realize that it is bad, then you have ruined the whole mixture. If you use two bowls, a small one in which you inspect the contents of a newly broken egg before adding it to the other eggs, and a large one containing all the (good) broken eggs, then you avoid ruining the mixture if the egg is bad. Now, however, you have two bowls to wash instead of one when you are finished. How would you reasonably define the loss in this case? While this example is perhaps not terribly serious, it illustrates one of the problems associated with trying to apply the Bayes strategy. Another problem is that it may be difficult to estimate prior probabilities for rare events. For example, suppose a country wants to use its antimissile resources against an attack by a hostile neighbor. If the defense strategy is designed according to a Bayes criterion, knowledge of the prior probability of an enemy attack is needed. How would one estimate this in a reasonable way?

7.2 LIKELIHOOD RATIO TEST

Because prior probabilities are often not available and loss functions may not be easily defined, we drop the constraint on minimizing the expected risk and modify the Bayes decision rule as

$$\begin{aligned}&\frac{f(X_1;\zeta_1)\cdots f(X_n;\zeta_1)}{f(X_1;\zeta_2),\dots f(X_n;\zeta_2)} > k, \text{ accept } \zeta_1 \text{ as state of nature}\\ &\frac{f(X_1;\zeta_1)\cdots f(X_n;\zeta_1)}{f(X_1;\zeta_2),\dots f(X_n;\zeta_2)} < k, \text{ reject } \zeta_1 \text{ as state of nature,}\end{aligned} \tag{7.2-1}$$

where the threshold value k is determined from criteria possibly other than that of Bayes. Common criteria are related to the probabilities of rejecting a claim when the claim is true and/or accepting the counterclaim when the counterclaim is true. This kind of test is known as a likelihood ratio test that tests a *simple hypothesis* (the claim) *against a simple alternative* (the counterclaim). To save on notation we define the LRT random variable as

$$\begin{aligned}\Lambda &\triangleq f(X_1;\zeta_1)\cdots f(X_n;\zeta_1)/f(X_1;\zeta_2)\cdots f(X_n;\zeta_2)\\ &= L(\zeta_1)/L(\zeta_2)\end{aligned} \tag{7.2-2}$$

and illustrate its application in an example.

Example 7.2-1

(*testing a claim for a food*) Consider a health-food manufacturer who claims to have developed a snack bar for kids that will reduce childhood obesity.[†] The snack bar, while tasty, supposedly acts as an appetite suppressant and thereby helps reduce the desire for fattening in-between-meals snacks such as potato chips, hamburgers, sugar-sweetened soda, chocolate bars, etc. To test the validity of this claim, we take n children (a subset of a large, well-defined group) and give them the weight-controlling snack bar. After one month, the average weight for this group is 98 lbs with a standard deviation of 5 lbs. The other children in the group, that is, the ones not taking the weight-controlling snack bar, average 102 lbs with a standard deviation of 5 lbs. We make the *hypothesis*[‡] that the weight-controlling snack bar has no effect in controlling obesity. This is called the *null hypothesis* and is denoted by H_1. The *alternative*, denoted by H_2, is that the weight-controlling snack is helpful in controlling obesity.[§] It does not matter which hypothesis we designate as H_1 but once the

[†]Obesity among children is a severe problem in the United States. Extrapolated from the present rate of caloric consumption, it is predicted that in 2020 three out of four Americans will be overweight or obese (Consumers Reports, December 2010, p. 11).

[‡]The meaning of this word is: an assumption provisionally accepted but currently not rigorously supported by evidence. A hypothesis is rejected if subsequent information doesn't support it.

[§]In many books the null hypothesis is denoted by H_0 and the alternative is denoted by H_a. We prefer the numerical subscript notation.

choice is made, we are required to be consistent throughout the problem. In the absence of well-defined loss function, we focus instead on the probabilities of error. We define

$$\alpha \triangleq P[\text{based on our test we decide that } H_2 \text{ is true}|H_1 \text{ is true}]$$
$$\beta \triangleq P[\text{based on our test we decide that } H_1 \text{ is true}|H_2 \text{ is true}].$$

With X_i, $i = 1, \ldots, n$, being i.i.d. RVs denoting the weights of the n children, we assume that the weights of both groups are Normally distributed near their means,† that is,

$$f(x_i, H_1) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - 102}{5}\right)^2\right]$$
$$f(x_i, H_2) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - 98}{5}\right)^2\right].$$

We note that $f_X(x_i, H_1)$ $(f_X(x_i, H_2))$ is the pdf of X_i, $i = 1, \ldots, n$, under the condition that $H_1(H_2)$ applies.

Then from Equation 7.2-2,

$$\Lambda = \exp\left(\frac{1}{2}\sum_{i=1}^{n}\left[\left(\frac{X_i - 98}{5}\right)^2 - \left(\frac{X_i - 102}{5}\right)^2\right]\right).$$

Further simplification yields

$$\Lambda = K_n \exp\left(\frac{4n}{25}\hat{\mu}_X(n)\right),$$

where $\hat{\mu}_X(n) \triangleq n^{-1}\sum_{i=1}^{n} X_i$ and K_n is a constant independent of the $\{X_i, i = 1, \ldots, n\}$ but dependent on the sample size n. The decision function then becomes

$$\text{if } K_n \exp\left(\frac{4n}{25}\hat{\mu}_X(n)\right) > k_n, \text{ accept } H_1, (\text{reject } H_2)$$
$$\text{if } K_n \exp\left(\frac{4n}{25}\hat{\mu}_X(n)\right) < k_n, \text{ accept } H_2, (\text{reject } H_1).$$

Since the natural logarithm of $\Lambda(\ln\Lambda)$ is an increasing function of Λ (Figure 7.2-1), we can simplify the decision function using (natural) logs and aggregating various constants into a single one. Then the test becomes

$$\text{if } \hat{\mu}_X(n) > c_n, \text{ accept } H_1$$
$$\text{if } \hat{\mu}_X(n) < c_n, \text{ accept } H_2,$$

where c_n is another constant that depends on the number of children in the test n, and is determined by the criterion we impose. If H_1 is true then $\hat{\mu}_X(n)$ is $N(102, 25/n)$ that is,

†The Normal characteristic is taken to be valid around the center of the pdf, say, a few σ values on either side of the mean. It definitely is not valid in the tails. For example, what would you make of a "negative weight"?

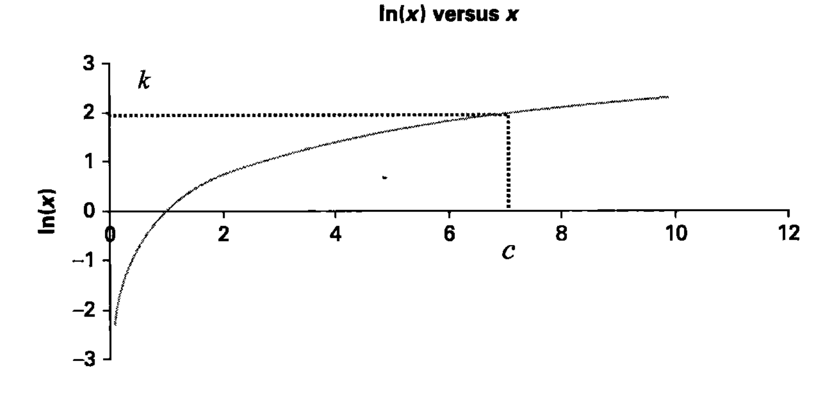


Figure 7.2-1 The natural logarithm of x is an increasing function of x.

$$f_{\hat{\mu}}(x, H_1) = \frac{1}{\sqrt{50\pi/n}} \exp\left(-\frac{1}{2}\left[\frac{x-102}{5/\sqrt{n}}\right]^2\right),$$

while if H_2 is true $\hat{\mu}_X(n)$ is $N(98, 25/n)$ that is,

$$f_{\hat{\mu}}(x, H_2) = \frac{1}{\sqrt{50\pi/n}} \exp\left(-\frac{1}{2}\left[\frac{x-98}{5/\sqrt{n}}\right]^2\right).$$

The pdf's of $\hat{\mu}_X(n)$ under H_1 and H_2 are shown in Figure 7.2-2.

Suppose by way of a criterion we specify $\alpha = 0.025$. Recall that $\alpha \triangleq P[\text{accept that } H_2 \text{ is true} | H_1 \text{ is true}]$. Then

$$0.025 = \int_{-\infty}^{c_n} f_{\hat{\mu}}(x, H_1)\,dx = F_{\hat{\mu}}(c_n) = F_{SN}\left(\frac{c_n - 102}{5/\sqrt{n}}\right) = F_{SN}(z_{0.025}),$$

which, from the Normal tables and simplifying, gives a threshold value $c_n = 102 - (9.8/\sqrt{n})$. As elsewhere the symbol $F_{SN}(z)$ stands for the CDF of the standard Normal RV evaluated at z. Thus, acceptance of H_1 requires the event $\{102 - (9.8/\sqrt{n}) < \hat{\mu}_X(n) < \infty\}$. This can also be written as $(102 - (9.8/\sqrt{n}), \infty)$ since intervals on the real line are events under RV mappings. The influence of the sample size on the threshold is shown in Figure 7.2-3. The power of the test increases with increasing sample size, as shown in (Figure 7.2-4). Increasing power means that the probability of making an error when H_2 is true decreases.

The error probability α is called the *probability of a type I error* and the *significance level* of *the test*.† The probability $\mathrm{P} \triangleq 1 - \beta$ is called the *power of the test* and β itself is called the *probability of a type II error*. The power of the test is the probability that we

†The error probability α is sometimes called the *size* of the test.

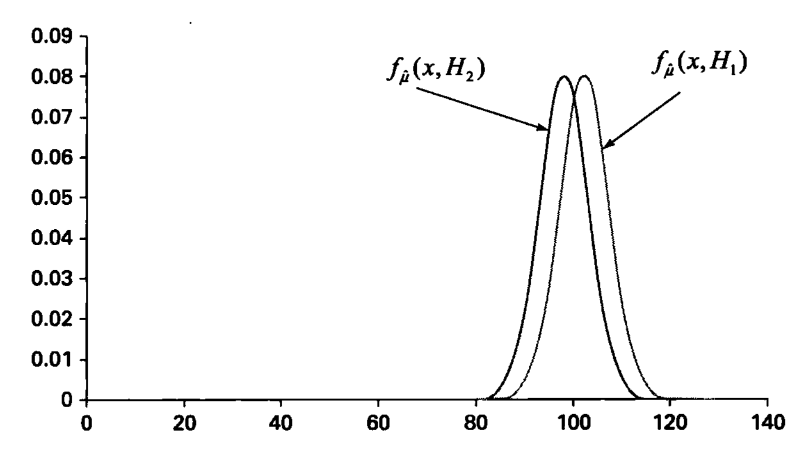


Figure 7.2-2 The pdf's $f_{\hat{\mu}}(x, H_1)$ and $f_{\hat{\mu}}(x, H_2)$ for Example 7.2-1.

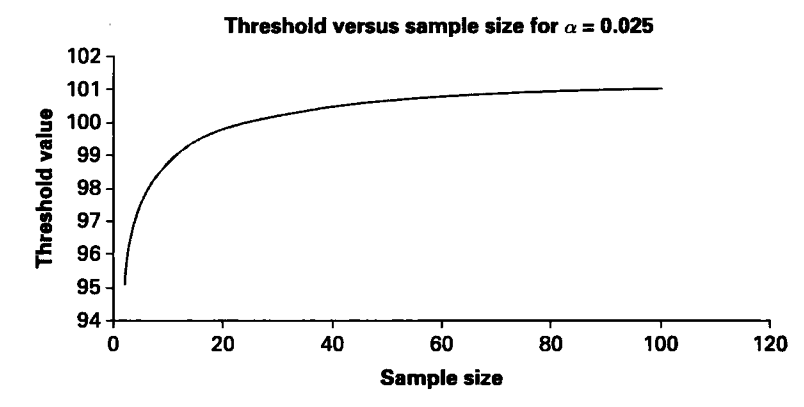


Figure 7.2-3 As the sample size increases, the threshold value moves to the right.

reject the null hypothesis given that the alternative is true. In general, it is not possible to make both α and β extremely small even *though it is not true*, in general, that $\alpha + \beta = 1$.

With reference to Example 7.2-1 we address a question some readers might have regarding this discussion, namely since the children eating the weight-controlling snack bar average 4 lbs less weight than their counterparts, why not simply accept this as evidence that the weight-controlling snack bar works? This would ignore the fact that even in the heavier group of children, a weight of 98 lbs is within one standard deviation from the mean of 102 lbs, meaning that if the sample size is small we could be in error in concluding that the snack bar is useful. Moreover, such a naïve approach would tell us nothing about the probability that we are mistaken.

Example 7.2-2

(*difference of means of Normal populations*) In some nutritional circles, there is a belief that bringing aid to Third-World malnourished children by way of a diet rich in omega-3 oils (e.g.,

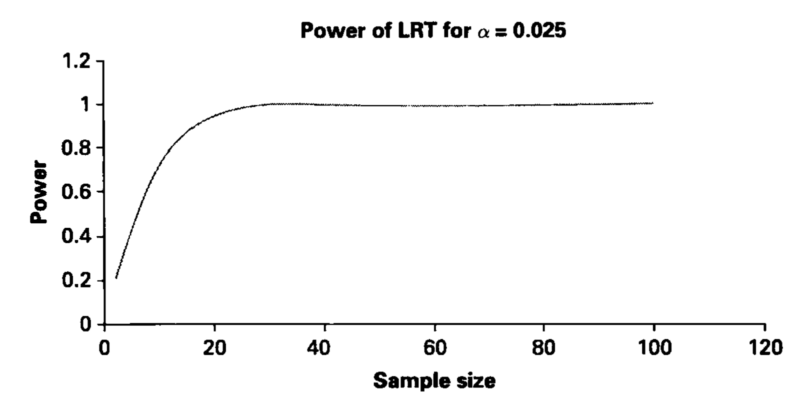


Figure 7.2-4 The power of the test increases with increasing sample size, which is a good thing. The best test would maximize the power of the test for a given n and α. In this example, the test is indeed the best test.

fish) and complex carbohydrates (whole wheat, bran, brown rice, etc.) can increase a child's IQ by 10 points by age 13, besides improving health. To test such a claim, one might want to measure the IQ of children brought up on such a diet against the IQ of children brought up on the local diet. Typically the data would be the sample mean $\hat{\mu}(n)$ of the IQs of the n children fed the experimental diet. If we denote the true but unknown mean by μ the test might take the form $H_1 : \mu_{\text{IQ}} = 110$ versus $H_2 : \mu_{\text{IQ}} = 100$. There are several variations on this type of test, for example, $H_1 : \mu = a$ versus $H_2 : \mu \neq a$ and $H_1 : a < \mu < b$ versus $H_2 : \mu < a, \mu > b$. We consider the elementary test $H_1 : \mu = b$ versus $H_2 : \mu = a(b > a)$ for a Normal population with, say, variance σ^2. We assume a random sample of size n, meaning that we have n i.i.d. RVs $X_1, \ldots, X_n$. Then if H_1 is true $X_i{:}N(b, \sigma^2)$ while if H_2 is true $X_i{:}N(a, \sigma^2)$. The LRT random variable is

$$\Lambda = \frac{\prod\limits_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2}\left[\frac{X_i - b}{\sigma}\right]^2\right)}{\prod\limits_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2}\left[\frac{X_i - a}{\sigma}\right]^2\right)}, \tag{7.2-3}$$

which, after simplifying, taking logs, and aggregating constants, yields the test

$$\begin{aligned} &\hat{\mu}(n) > c_n, \text{ accept } H_1 (\text{reject } H_1) \\ &\hat{\mu}(n) < c_n, \text{ reject } H_1 \text{ (accept } H_2). \end{aligned}$$

The constant c_n is determined by our choice of α. For example with $\alpha = 0.025$, we must solve

$$\begin{aligned} \alpha &= P[\text{accept } H_2 | H_1 \text{ true}] = 0.025 \\ &= \int_{-\infty}^{c_n} \frac{1}{(2\pi)^{0.5}\sigma/\sqrt{n}} \exp\left(-1/2\left[\frac{y-b}{\sigma/\sqrt{n}}\right]^2\right) dy \\ &= \int_{-\infty}^{c_n'} 1/(2\pi)^{0.5} \exp\left(-1/2y^2\right) dy = F_{SN}\left(\frac{c_n - b}{\sigma/\sqrt{n}}\right) = F_{SN}(z_{0.025}), \end{aligned}$$

where $c'_n \triangleq (c_n - b)\sqrt{n}/\sigma$. From the tables of the standard Normal CDF, we find that $z_{0.025} = -1.96$. Solving, we get $c_n = b - (1.96\sigma/\sqrt{n})$. Notice the similarity between this example and Example 7.2-1. The power of the test is

$$\begin{aligned} \text{P} &= 1 - P[\text{accept } H_1|H_2 \text{ true}] \\ &= 1 - (2\pi\sigma^2/n)^{-1/2} \int_{c_n}^{\infty} \exp\left(-0.5\left[\frac{z-a}{\sigma/\sqrt{n}}\right]^2\right) dz \qquad (7.2\text{-}4) \\ &= F_{SN}\left(\frac{b-a-(1.96\sigma/\sqrt{n})}{\sigma/\sqrt{n}}\right). \end{aligned}$$

The reader will recognize that the power of the test is simply the probability of accepting H_2 when H_2 is true. Returning to the IQ problem that motivated this discussion, we find that for $\alpha = 0.025$, $b = 110, a = 100$, $\sigma = 10$, and $n = 25$, the acceptance region for H_1 is the region to the right of $c_n = 106.1$. In other words, when the event $\{106.1 < \hat{\mu}(n) < \infty\}$ occurs, it suggests that a good diet helps to overcome the IQ deficiency of malnourished children. The power of the test is 0.999.

Neyman–Pearson Theorem. Suppose we are asked to find a test for a simple hypothesis versus a simple alternative that, for a given α, will minimize β. Such a test will maximize the power $\text{P} = 1 - \beta$ and is therefore *a most powerful test.* What is this test? The Neyman–Pearson theorem (given here without proof) furnishes the answer.

Theorem 7.2-1 Denote the set of points in the critical region by R_k (i.e., the region of outcomes where we reject the hypothesis H_1). Denote the *significance of the test* as α meaning $P[\text{accept } H_2|H_1 \text{ is true}] \leq \alpha$. Then R_k maximizes the power of the test $\text{P} \triangleq 1 - \beta$ if it satisfies

$$\Lambda \triangleq \frac{f(X_1, \zeta_1)\cdots f(X_n;\zeta_1)}{f(X_1, \zeta_2)\cdots f(X_n;\zeta_2)} < k \qquad (7.2\text{-}5)$$

for some fixed number k, which determines R_k. ■

Discussion. The Neyman–Pearson Theorem (NPT) says that the likelihood ratio test, subject to the above constraints, that is, at significance α, is the *most powerful test.* In this sense it is an optimal test. The relationship between R_k, k, and α is not explicitly stated by the theorem but becomes clear in working a problem.

Example 7.2-3

(*chicken feed for making large eggs*) A producer of chicken feed claims that a new product "Eggrow," when fed to chickens, will cause the laid eggs to be larger than those laid by chickens fed ordinary feed. With ordinary feed, the chickens raised by this producer lay eggs that on the average weigh 60 grams per egg, with a standard deviation of 4 grams. Twenty-five chickens fed on "Eggrow" produce eggs whose average weight is 62 grams with a standard deviation of 4 grams. Let the hypothesis be $H_1 : \mu = \mu_1 = 62$ and the alternative be $H_2 : \mu = \mu_1 = 60$. The significance level of the test is 0.05. According to the NPT, the test

$$\Lambda = \frac{\prod_{i=1}^{n} (2\pi 16)^{-1/2} \exp\left(-\frac{1}{2}\left[\frac{X_i - 62}{4}\right]^2\right)}{\prod_{i=1}^{n} (2\pi 16)^{-1/2} \exp\left(-\frac{1}{2}\left[\frac{X_i - 60}{4}\right]^2\right)} < k$$

that defines the critical region R_k is the most powerful test. Then $\Lambda = \exp\left(\frac{n\hat{\mu}}{8} + (60)^2 - (62)^2\right)$ and taking logs, aggregating constants, and simplifying, yields the test

$$\begin{aligned} &\text{if } \hat{\mu} < c_n, \text{ reject } H_1, \text{ accept } H_2 \\ &\text{if } \hat{\mu} > c_n, \text{ accept } H_1, \text{reject } H_2, \end{aligned}$$

where c_n is an unknown constant. To find c_n and the rejection region R_k, we solve

$$0.05 = \int_{-\infty}^{c_n} \frac{1}{\sqrt{2\pi}(4/5)} \exp\left(-\frac{1}{2}\left(\frac{z-62}{0.8}\right)^2\right) dz$$

and find that $c_n = 60.7$ and $R_k = (0, 60.7)$. Thus, if $\hat{\mu} < 60.7$, reject H_1, accept H_2. The test is most powerful and $\text{P} \approx 0.81$.

7.3 COMPOSITE HYPOTHESES

In the previous section we mentioned that in practice there are tests of the form: $H_1 : a < \mu < b$ versus $H_2 : \mu < a, \mu > b$ and others. All of these tests have one thing in common: either H_1 or H_2 or both deal with events whose sample space has many outcomes. In the case of the simple hypothesis versus the simple alternative, the sample space had only two points ζ_1 and ζ_2. In the case of composite hypotheses, the test

$$\Lambda \triangleq \frac{f(X_1, \zeta_1) \cdots f(X_n; \zeta_1)}{f(X_1, \zeta_2) \cdots f(X_n; \zeta_2)} < k \tag{7.3-1}$$

has no meaning because there are many more ζ's than just ζ_1 and ζ_2. To understand the material in this section, the reader should recall that in the estimation of parameters by the *maximum likelihood method* (MLM) the idea was to find the parameter θ in the likelihood function $L(\theta)$ that was *most likely* to have yielded the observed result. Often this could be found by differentiation but not always. In the problems discussed so far, there was no need to maximize the likelihood function to find the most likely θ because θ had one of two values, either ζ_1 or ζ_2. Suppose that the parameter of interest is the mean, that is, $\theta = \mu$. Then in a problem such as $H_1 : \mu = \mu_0$ versus $H_2 : \mu \neq \mu_0$ the maximization of the likelihood function associated with H_2 requires *searching* for the optimum value of μ in the *parameter space* $(-\infty, \infty)$. In other words, while the hypothesis in this case is simple, the alternative is not: It is said to be *composite*.

Fortunately not all composite hypothesis problems require such a search. We can still use the Neyman–Pearson rule and its desirable *most-powerful* property. We illustrate with an example.

Example 7.3-1

(*testing the hypothesis* $\mathrm{H}_1 : \mu = \mu_1$ *versus the alternative* $\mathrm{H}_2 : \mu < \mu_1$) We assume a Normal population with mean μ and variance σ^2. At first glance it would seem that the likelihood function associated with $H_2 : \mu < \mu_1$ requires a search. However, we can reduce this problem to a simple hypothesis versus a simple alternative by a slight modification of the H_2 hypothesis. That is, we modify the problem to $H_1 : \mu = \mu_1$ versus $H_2' : \mu = \mu_2 < \mu_1$, where μ_2 is as yet arbitrary. Then

$$\Lambda = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i-\mu_1)^2 - \sum_{i=1}^{n}(X_i-\mu_2)^2\right)\right) < k \qquad (7.3\text{-}2)$$

is the LRT for the critical region for H_1. Simplifying, taking logs, and aggregating all constants, we obtain the test: if $\hat{\mu} < c_n$ reject H_1. To find the constant c_n we proceed as before; that is, we use the type I error criterion, that is, the *significance level* of the test. Thus, say, with $\alpha = 0.01$ and the pdf $f_{\hat{\mu}}(z; \mu_1) = N(\mu_1, \sigma^2/n)$ we solve

$$0.01 = \frac{\sqrt{n}}{\sqrt{2\pi}}\int_{-\infty}^{c_n} \exp\left[-0.5\left(\frac{z-\mu_1}{\sigma/\sqrt{n}}\right)^2\right] dz,$$

to obtain $c_n = \mu_1 - 2.32\sigma/\sqrt{n}$. Thus, we reject H_1 if $\hat{\mu} < \mu_1 - 2.32\sigma/\sqrt{n}$. Note that we never had to specify an actual value for μ_2.

Generalized Likelihood Ratio Test (GLRT)

The GLRT is useful for solving *composite hypotheses problems*. First, recall that some likelihood functions are functions of one parameter, some of two parameters, etc. For example, the likelihood function associated with an n-sample of i.i.d. *exponential* RVs is $L(\lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} X_i)u(X_i)^\dagger$ and is a function only of the parameter $\theta = \lambda$, while the likelihood function associated with an n-sample of i.i.d. Normal RVs is

$$L(\mu, \sigma) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{X_i - \mu}{\sigma}\right]^2\right)$$

and is a function of two parameters $\boldsymbol{\theta} = (\mu, \sigma)$. The likelihood function associated with a *two-dimensional* (*multivariate*) Normal would be a function of five parameters, that is, $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}$. We use the notation $L(\boldsymbol{\theta})$ to indicate a likelihood function of the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Now consider the following problem: Let Θ denote the global k-dimensional parameter space; for example, in the univariate Normal this would be $\Theta = (-\infty < \mu < \infty, 0 < \sigma < \infty)$. Let Θ_1 denote the parameter space (a subspace of Θ) associated with the hypothesis H_1. For example if $X{:}N(\mu_X, \sigma_X^2)$ and the hypothesis

†The function $u(x)$ is the unit step: $u(x) = 1, x \geq 0$, and zero elsewhere.

is H_1: $3 < \mu_X < 4$, then $\Theta_1 = (3 < \mu_X < 4,\ 0 < \sigma_X^2 < \infty)$. Define the test statistic Λ for testing $H_1 : \boldsymbol{\theta} \in \Theta_1$ versus the alternative $H_2 : \boldsymbol{\theta} \notin \Theta_1$ as

$$\Lambda \triangleq \frac{L_{LM}(\boldsymbol{\theta}^*)}{L_{GM}(\boldsymbol{\theta}^\dagger)}, \tag{7.3-3a}$$

where $L_{LM}(\boldsymbol{\theta}^*) \triangleq \max_{\boldsymbol{\theta}\in\Theta_1} L(\boldsymbol{\theta})$ and $L_{GM}(\boldsymbol{\theta}^\dagger) \triangleq \max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$. We may ask why Λ, as given in Equation 7.3-3a, is a reasonable test statistic for accepting or rejecting H_1. First recall that maximizing the numerator gives us the most likely parameter estimates, restricted to Θ_1, to account for the observations. Because our search is restricted to Θ_1, the maximum in this parameter subspace may not be a global maximum; hence we call it a *local maximum.* Next, maximizing the denominator gives us the most likely *unrestricted* parameter estimates that account for the observations; hence we call it a *global maximum.* The subscripts LM and GM are there to remind the reader of the "local-max" and "global-max" operations, respectively. We observe that Λ is a random variable with its realization confined to [0,1]. (Question for the reader: Why is this so?). Now if the realizations of Λ are close to one, then we assume that H_1 is true; that is, the unknown parameters are in Θ_1 but, in fact, are also the most likely parameters in the whole space. On the other hand, if the realizations of Λ are small or close to zero we may assume that the most likely parameters are not in Θ_1. The *threshold value c* denotes the point at which we go from accepting (rejecting) the hypothesis to accepting (rejecting) the alternative H_2 and is usually determined by enforcing the significance level α. In summary then, the GLRT is described as

$$\text{reject } H_1 \text{ if } \Lambda < c, \tag{7.3-3b}$$

where Λ is given in Equation 7.3-3a. It has been shown that under certain conditions, the GLRT is asymptotically optimal in the Neyman-Pearson sense. However there exist counter-examples in the literature that prove that the GLRT is not always optimal [7-22]. In this sense it must be regarded as being *empirical.*

We illustrate the application of the GLRT with several examples involving continuous distributions.

Example 7.3-2

(*testing* $H_1 : \mu = \mu_1$ *versus* $H_2 : \mu \neq \mu_1$ *when* X *is Normal and* σ^2 *known*) We make n observations on a Normal RV with *known* variance σ^2. The likelihood function is

$$\begin{aligned} L(\mu) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(X_i-\hat{\mu})^2+(\hat{\mu}-\mu)^2\right]\right). \end{aligned} \tag{7.3-4}$$

To go from line 1 to line 2 we generated some cross-terms in the argument of the exponent that vanish in the summation. We leave the algebraic steps as an exercise to the reader. Since σ^2 is specified, the space Θ is $(-\infty < \mu < \infty)$. Then $L_{GM}(\mu^\dagger)$ is obtained when $\mu^\dagger = \hat{\mu}$, that is, $L_{GM}(\mu^\dagger) = L(\hat{\mu})$, and since Θ_1 contains only one point it follows that $L_{LM}(\mu^*) = L(\mu_1)$. From Equation 7.3-3a we get

$$\Lambda = \frac{L(\mu_1)}{L(\hat{\mu})} = \exp\left(-\frac{n}{2\sigma^2}(\hat{\mu}-\mu_1)^2\right) \tag{7.3-5}$$

and the critical region is associated with outcomes of $\hat{\mu}$ that are far from μ_1. When $\hat{\mu}$ is near μ_1, Λ will take values near 1 and we would tend to accept H_1. Likewise when $\hat{\mu}$ is far from μ_1 it is unlikely that H_1 is true and we reject it. Somewhere in between is a constant c such that $0 < \Lambda < c$ describes the critical region. Taking natural logs, we find that the critical region is defined by

$$\hat{\mu} > \mu_1 + \left(2\sigma^2 \ln(1/c)/n\right)^{1/2}$$
$$\hat{\mu} < \mu_1 - \left(2\sigma^2 \ln(1/c)/n\right)^{1/2}, \tag{7.3-6}$$

where c is determined by the significance level α of the test.

Example 7.3-3

(*numerical realization in Example 7.3-2*) Here we obtain a numerical evaluation Equation 7.3-6. Assume that $\mu_1 = 5$, $\sigma^2 = 4, n = 15$, and $\alpha = 0.05$. With $f_\Lambda(x;\mu_1)$ denoting the pdf of Λ we must compute $0.05 = \int_0^c f_\Lambda(x)dx$. But the event $\{\Lambda \leq c\}$ is identical to the event$\{-\infty < \ln\Lambda < \ln c\}$, which in turn is identical to $\{-2\ln c \leq -2\ln\Lambda < \infty\}$. From Equation 7.3-5, $-2\ln\Lambda = \left(\frac{\hat{\mu}-\mu_1}{\sigma/\sqrt{n}}\right)^2$, which is χ^2 with one degree of freedom, that is, χ_1^2 (the subscript indicates the degree of freedom). Denoting the χ_n^2 pdf by $f_{\chi^2}(x;n)$ we write

$$0.05 = \int_0^c f_\Lambda(x)dx = \int_{-2\log c}^{\infty} f_{\chi^2}(x;1)dx = 1 - F_{\chi^2}(-2\ln c;1).$$

From the tables of the CDF of the χ_1^2 RV we obtain $-2\ln c = 3.84$. Hence from Equation 7.3-6 we determine the critical region as $\hat{\mu} > 6.01, \hat{\mu} < 3.99$ or, as interval events mapped by $\hat{\mu}$ $(-\infty, 3.99) \cup (6.01, \infty)$.

Example 7.3-4

(*testing the telephone waiting time when the call is in a queue*) A call to the Goldmad Investment Bank (GIB) gets an automatic (robotic) operator that announces that during business hours the average waiting time to speak to an investment consultant is less than 30 seconds (0.5 minutes). We wish to test this claim using the GLRT. We make n calls to the GIB during business hours and record the waiting times $X_i, i = 1, \ldots, X_n$, assumed to be i.i.d. exponential random variables each with pdf $f_{X_i}(x;\mu) = (1/\mu)\exp(-x/\mu)\,u(x)$, where $\mu = E(X_i), i = 1, \ldots, n$. From basic probability we know that $\hat{\mu} = (1/n)\sum_{i=1}^{n} X_i$ is an unbiased, consistent estimator for μ. We test the hypothesis $H_1 : \mu \leq 0.5$ versus $H_2 : \mu > 0.5$. The likelihood function is $L(\mu) = (1/\mu)^n \exp\left(-\frac{n}{\mu}\left[(1/n)\sum_{i=1}^{n} X_i\right]\right) = (1/\mu^n)\exp(-n\hat{\mu}/\mu)$. Then $L_{GM}(\mu^\dagger)$ is obtained by differentiation with respect to μ to obtain

$$L_{GM}(\mu^\dagger) = L(\hat{\mu}) = \hat{\mu}^{-n}\exp(-n).$$

Finding $L_{LM}(\mu^*)$ is a little more sophisticated. To illustrate what is going on we plot two likelihood functions in Figure 7.3-1, one that peaks at the mean of 0.45 and another that peaks at the mean of 0.55. The μ space $\Theta_1 = (0, 0.5]$ is based on our hypothesis that $\mu \leq 0.5$ and includes the global maximum point 0.45. However, when the likelihood function is the

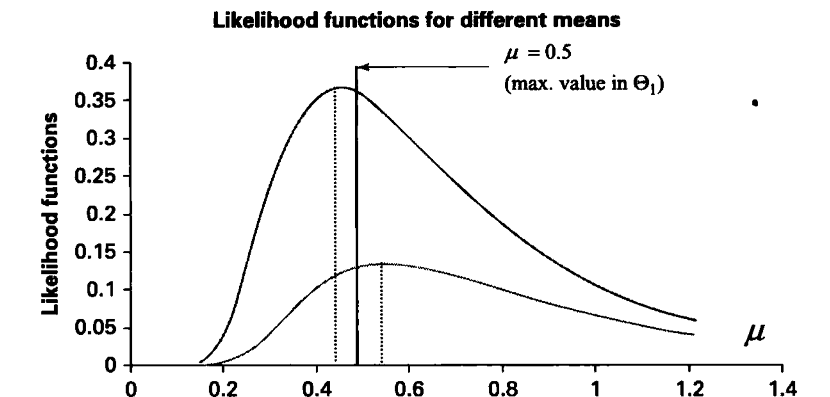


Figure 7.3-1 The upper curve is the likelihood function when the true mean is at 0.45 and $n = 10$. The lower curve is the likelihood function when the true mean is at 0.55 and $n = 10$. The subspace $\Theta_1 = (0,0.5]$ includes the point 0.45 (shown as the dotted line on the left of the solid line) but not the point 0.55 (shown as the dotted line on the right of the solid line).

lower curve in Figure 7.3-1, which peaks at $\mu = 0.55$, the local maximum is not the same as the global maximum since the point 0.55 is not in $\Theta_1 = (0, 0.5]$.

Hence

$$L_{LM}(\mu^\dagger) = \begin{cases} \hat{\mu}^{-n} \exp(-n), \hat{\mu} \leq 0.5 \\ 2^n \exp(-2n\hat{\mu}), \hat{\mu} > 0.5. \end{cases}$$

The subspace $\Theta_1 = (0, 0.5]$ includes the point 0.45 (shown as the dotted line on the left of the solid line) but not the point 0.55 (shown as the dotted line on the right of the solid line).

Finally, from Equation 7.3-3a, we get

$$\begin{aligned} \Lambda &\triangleq \frac{L_{LM}(\mu^*)}{L_{GM}(\mu^\dagger)} \\ &= \begin{cases} 1, \ \hat{\mu} \leq 0.5 \\ (2\hat{\mu})^n \exp\left(-n[2\hat{\mu} - 1]\right), \ \hat{\mu} > 0.5. \end{cases} \end{aligned} \tag{7.3-7}$$

The critical region is the interval $(0, c')$; that is, all outcomes $\Lambda \in (0, c')$ would lead to the rejection of H_1. The critical region is shown in Figure 7.3-2: On the Λ axis it is below the horizontal line at c'; on the $\hat{\mu}$ axis it is to the right of $\hat{\mu} = c$.

Because the likelihood function decreases monotonically with $\hat{\mu}$ in the region $\hat{\mu} > 0.5$ (Figure 7.3-2), we can use $\hat{\mu}$ as a test statistic. Assuming that n is large enough for the Normal approximation to apply to the behavior of $\hat{\mu}$, at least where the pdf has significant value, that is, within a few sigmas around its mean, we have $\hat{\mu} : N(\mu, \mu^2/n)$ since $\hat{\mu}$ is an unbiased estimator for μ. In writing this result we recalled that the variance of a single exponential RV is μ^2 and therefore the variance of $\widehat{\mu}$ is $\sigma^2_{\hat{\mu}} = \mu^2/n$. We create the approximate standard Normal RV from $Z \triangleq (\hat{\mu} - \mu)\sqrt{n}/\mu$ and compute c from the

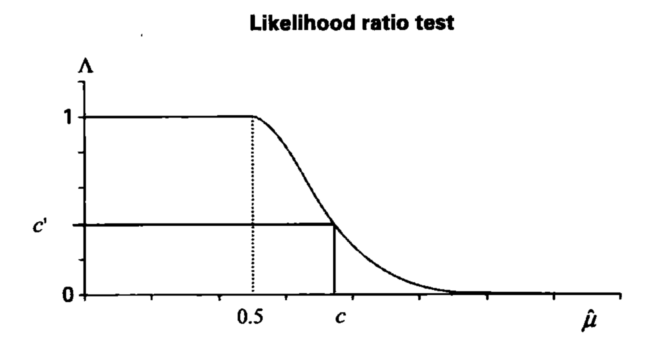


Figure 7.3-2 Variation of the GLR test statistic with the sample mean estimator.

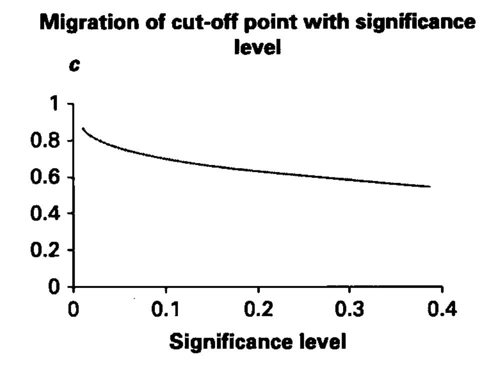


Figure 7.3-3 As α increases, the cut-off point decreases thereby increasing the width of the critical region.

significance constraint α. Using the percentile notation $1 - \alpha = F_{SN}(z_{1-\alpha})$, we find that $c = \mu + z_{1-\alpha}\mu/\sqrt{n}$, from which we see that the critical point c increases linearly with μ. We reject the hypothesis when $\hat{\mu} > c$. For example with $\mu = 0.5, \alpha = 0.05$, and $n = 10$ we find that $z_{0.95} = 1.64$ and $c \approx 0.76$. As α increases, the cut-off point decreases toward 0.5 (Figure 7.3-3).

Example 7.3-5

(*evaluation of cancer treatment by the drug Herceptin*) Newer treatments for cancer involve disabling the proteins that fuel cancer. For example, some breast cancers contain a protein called HER2. In such cases, the drug Herceptin is partially effective in treating the cancer in that it reduces the cancer recurrence by 50 percent.† Tumors that do not exhibit HER2 have better prognoses than those that do. Since Herceptin has significant toxic side effects, it is important that the test for the HER2 protein is accurate but this is not always the case. Let H_1: tumor has a high level of HER2 and therefore will respond to Herceptin, and

† "Cancer Fight: Unclear Tests for New Drug," *New York Times*, April 20, 2010.

let H_2: tumor has low levels (or none at all) of HER2 and therefore the patient should not be given Herceptin. It is estimated that in current testing for HER2:

$$P[\text{decide } H_1 \text{ is true}|H_2 \text{ is true}] = 0.2$$
$$P[\text{decide } H_2 \text{ is true}|H_1 \text{ is true}] = 0.1.$$

Hence, the tests have a significance level of 0.1 and a power of 0.8.

How Do We Test for the Equality of Means of Two Populations?

Assume that there is a drug being tested for androgen-independent prostate cancer.† The drug is administered to a group of men with advanced prostate cancer. Does the drug extend the lives of the participants compared with those of men taking the traditional therapy? A printing company is evaluating two types of paper for use in its presses. Is one type of paper less likely to jam the presses than the other? The Department of Transportation is considering buying concrete from two different sources. Is one more resistant to potholes than the other is? Some of these problems fall within the following framework. We have two populations, assumed Normal, and we have m samples from population P1 and n samples from population P2. Is the mean of population P1 equal to the mean of population P2? In general, this is a difficult problem, essentially beyond the scope of the discussion treated in this chapter. More discussion on this problem is given in [7-1]. However, when one can assume that the variance of the populations is the same, the problem is treatable analytically in a straightforward way. In preparation for discussing this problem, we review some related material in Example 7.3-6.

Example 7.3-6

(*preliminary results for Example 7.3-7*) We have samples from two Normal populations $S_1 = \{X_{1i}, i = 1, \ldots, m\}$ and $S_2 = \{X_{2i}, i = 1, \ldots, n\}$. The elements of S_1 are m i.i.d. observations on X_1 with $X_1{:}N(\mu_1, \sigma_1^2)$. Likewise, the elements of S_2 are n i.i.d. observations on X_2 with $X_2{:}N(\mu_2, \sigma_2^2)$. Further, assume that $E[(X_{1i} - \mu_1)(X_{2j} - \mu_2)] = 0$, all i, j.

(i) Assuming $\mu_1 = \mu_2 = \mu$, show that $E[\hat{\mu}_1 - \hat{\mu}_1] = 0$.

Solution to (i) $E[\hat{\mu}_1 - \hat{\mu}_2] = E[\hat{\mu}_1] - E[\hat{\mu}_2] = \mu - \mu = 0$.

(ii) Assume that $\mu_1 = \mu_2 = \mu$ and $\sigma_1 = \sigma_2 = \sigma$. Show that $\text{Var}(\hat{\mu}_1 - \hat{\mu}_2) = (m^{-1} + n^{-1})\sigma^2$.

Solution to (ii) Since $E[\hat{\mu}_1] = E[\hat{\mu}_2] = \mu$, $\text{Var}(\hat{\mu}_1 - \hat{\mu}_2) = E[(\hat{\mu}_1 - \hat{\mu}_2)^2] = E[\hat{\mu}_1^2] + E[\hat{\mu}_2^2] - 2E[\hat{\mu}_1\hat{\mu}_2]$. Substitute

$$E[\hat{\mu}_1^2] = m^{-2}\left(\sum_{i=1}^{m} E(X_{1i}^2) + \sum_{i=1}^{m}\sum_{j\neq i}^{m} E(X_{1i}X_{1j})\right)$$
$$E[\hat{\mu}_2^2] = n^{-2}\left(\sum_{i=1}^{n} E(X_{2i}^2) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} E(X_{2i}X_{2j})\right)$$
$$E(X_{1i}^2) = \mu^2 + \sigma^2 = E(X_{2i}^2) \text{ and}$$
$$E(\hat{\mu}_1\hat{\mu}_2) = \mu^2$$

into the expression for the variance and obtain the required result.

---

† Androgen-independent means that the cancer is not fueled by testosterone. It is difficult to treat. The authors' colleague, Prof. Nick Galatsanos, an important contributor to the science of image processing, died from this illness at the age of 52.

(iii) Show that if V and W are Chi-square with *degrees of freedom* (DOF) m and n, respectively, then $U \triangleq V + W$ is Chi-square with DOF $m + n$.

Solution to (iii) If $V : \chi^2_m$ and $W : \chi^2_n$ then $V = \sum_{i=1}^m Y_i^2$ and $W = \sum_{i=1}^n Z_i^2$, where we can assume that $Y_i, i = 1, \ldots, m$, are i.i.d. $N(0, 1)$ and Z_i, $i = 1, \ldots, n$, are i.i.d. $N(0, 1)$. The MGF of V is $M_V(t) = E[\exp(tV]$ and is computed as

$$M_V(t) = (2\pi)^{-m/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp(t \sum_{i=1}^m y_i^2) \times \exp(-1/2 \sum_{i=1}^m y_i^2) \prod_{i=1}^m dy_i$$
$$= \prod_{i=1}^m (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-0.5(1 - 2t)y_i^2\right) dy_i$$
$$= (1 - 2t)^{-m/2} \text{ for } t < 1/2.$$

Line 1 is by definition; line 2 is by the i.i.d. assumption on the Y_i's; and line 3 results from the total area under the Normal curve being unity. Because $U = V + W$ and V and W are jointly independent, it follows from the discussion in Section 4.4 that $M_U(t) = M_V(t)M_W(t)$. Since $M_V(t) = (1 - 2t)^{-m/2}$ and $M_W(t) = (1 - 2t)^{-n/2}$ it follows that $M_U(t) = (1 - 2t)^{-(m+n)/2}$, which implies that $U : \chi^2_{m+n}$.

(iv) Given the likelihood function $L = (2\pi\sigma^2)^{-m/2} \exp[-0.5 \sum_{i=1}^m ((X_i - \mu)^2/\sigma^2)]$ show that $L_{GM} = L(\hat{\mu}^\dagger, \hat{\sigma}^{2\dagger})$, where, in this case, $\hat{\mu}^\dagger = \hat{\mu}$ and $\hat{\sigma}^{2\dagger} = \hat{\sigma}^2$

Solution to (iv) We obtain $\hat{\mu}^\dagger$ by differentiating $\ln L$ with respect to μ and obtain $\hat{\mu}^\dagger = \hat{\mu} \triangleq (m)^{-1} \sum_{i=1}^m X_i$. Likewise, we obtain $\hat{\sigma}^{2\dagger}$ by differentiating with respect to σ^2 and obtain $\hat{\sigma}^{2\dagger} = \hat{\sigma}^2 \triangleq m^{-1} \sum_{i=1}^m (X_i - \hat{\mu})^2$. Substituting into the expression for L we compute L_{GM} as

$$L_{GM} = \left(\frac{m}{2\pi \sum_{i=1}^m (X_i - \hat{\mu})^2} \right)^{m/2} e^{-m/2}. \tag{7.3-8}$$

Example 7.3-7

(*testing* $H_1 : \mu_1 = \mu_2$ *versus* $H_2 : \mu_1 \neq \mu_2, \sigma_1^2 = \sigma_2^2 = \sigma^2$ *not known*) As in Example 7.3-6 we have samples from two Normal populations $S_1 = \{X_{1i}, i = 1, \ldots, m\}$ and $S_2 = \{X_{2i}, i = 1, \ldots, n\}$. The elements of S_1 are m i.i.d. observations on X_1 with $X_1 : N(\mu_1, \sigma_1^2)$. Likewise, the elements of S_2 are n i.i.d. observations on X_2 with $X_2 : N(\mu_2, \sigma_2^2)$. Further, assume that $E[(X_{1i} - \mu_1)(X_{2j} - \mu_2)] = 0$, for all i, j. We shall test $H_1 : \mu_1 = \mu_2$ versus $H_2 : \mu_1 \neq \mu_2$. The parameter space† for H_1 is $\Theta_1 = (\mu, \sigma^2)$ while the global parameter space is $\Theta = (\mu_1, \mu_2, \sigma^2)$. The likelihood function is

$$L = \left(\frac{1}{2\pi\sigma^2} \right)^{(m+n)/2} \exp\left(-\frac{1}{2} \sum_{i=1}^m \left(\frac{X_{1i} - \mu_1}{\sigma} \right)^2 \right) \times \exp\left(-\frac{1}{2} \sum_{i=1}^n \left(\frac{X_{2i} - \mu_2}{\sigma} \right)^2 \right) \tag{7.3-9}$$

†To avoid excessive notation we denote a parameter space such as $\Theta = \{-\infty < \mu_1 < \infty, -\infty < \mu_2 < \infty, \sigma_1 > 0, \sigma_2 > 0\}$ by $\Theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2\}$ etc. for other cases, when there is no danger of confusion. Then the expression $L(\Theta)$ can be interpreted as the likelihood function of parameters in the space Θ.

and from the results of (iv) in Example 7.3-6 we obtain

$$\hat{\mu}_1^\dagger = m^{-1}\sum_{i=1}^m X_{1i} = \hat{\mu}_1$$
$$\hat{\mu}_2^\dagger = n^{-1}\sum_{i=1}^n X_{2i} = \hat{\mu}_2$$
$$\hat{\sigma}^{2\dagger} = \tfrac{1}{m+n}\left(\sum_{i=1}^m (X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^n (X_{2i}-\hat{\mu}_2)^2\right) = \hat{\sigma}^2.$$

We insert these results in L for μ_1, μ_2, and σ^2, respectively, to obtain

$$L_{GM} = \left(\frac{m+n}{2\pi\left(\sum_{i=1}^m (X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^n (X_{2i}-\hat{\mu}_2)^2\right)}\right)^{(m+n)/2} \exp\left(-\frac{(m+n)}{2}\right). \tag{7.3-10}$$

Returning now to the likelihood function in Equation 7.3-9, we wish to maximize this in the parameter subspace Θ_1. Since in H_1 $\mu_1 = \mu_2 = \mu$, we rewrite L as

$$L(\mu,\sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{(m+n)/2} \exp\left(-\frac{1}{2}\sum_{i=1}^m \left(\frac{X_{1i}-\mu}{\sigma}\right)^2\right)\times\exp\left(-\frac{1}{2}\sum_{i=1}^n \left(\frac{X_{2i}-\mu}{\sigma}\right)^2\right). \tag{7.3-11}$$

Straightforward differentiation with respect μ and σ^2 yields $\hat{\mu}^*$ and $\hat{\sigma}^{2*}$ as

$$\begin{aligned}\hat{\mu}^* &= \frac{1}{m+n}\left(\sum_{i=1}^m X_{1i} + \sum_{i=1}^n X_{2i}\right)\\ &= \frac{m}{m+n}\hat{\mu}_1 + \frac{n}{m+n}\hat{\mu}_2\end{aligned}$$

and

$$\hat{\sigma}^{2*} = \frac{1}{m+n}\left(\sum_{i=1}^m (X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^n (X_{2i}-\hat{\mu}_2)^2 + \frac{mn}{m+n}(\hat{\mu}_1-\hat{\mu}_2)^2\right).$$

When $\hat{\mu}^*$ and $\hat{\sigma}^{2*}$ are substituted for μ and σ^2 in $L(\Theta_1)$ of Equation 7.3-11, we obtain L_{LM} as

$$L_{LM} = \left[\frac{(m+n)e^{-1}}{2\pi\left(\sum_{i=1}^m (X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^n (X_{2i}-\hat{\mu}_2)^2 + \frac{mn}{m+n}(\hat{\mu}_1-\hat{\mu}_2)^2\right)}\right]^{(m+n)/2}.$$

The likelihood ratio $\Lambda \triangleq L_{LM}/L_{GM}$ is computed as

$$\Lambda = \left[1 + \frac{\frac{mn}{m+n}(\hat{\mu}_1-\hat{\mu}_2)^2}{\sum_{i=1}^m (X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^n (X_{2i}-\hat{\mu}_2)^2}\right]^{-(m+n)/2}. \tag{7.3-12}$$

From (ii) in Example 7.3-6, $\hat{\mu}_1 - \hat{\mu}_2$ is distributed as $N\left(0, \sigma^2(m+n)/mn\right)$, so that

$$Z \triangleq \frac{(\hat{\mu}_1-\hat{\mu}_2)}{\sigma\left(\frac{m+n}{mn}\right)^{1/2}}$$

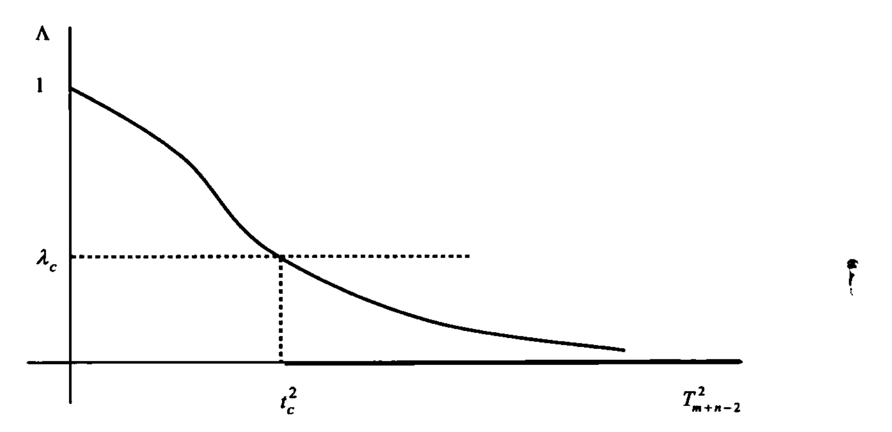


Figure 7.3-4 Critical region shown in heavy lines. It is easier to test H_1 versus H_2 using a test on T^2_{m+n-2} than on Λ.

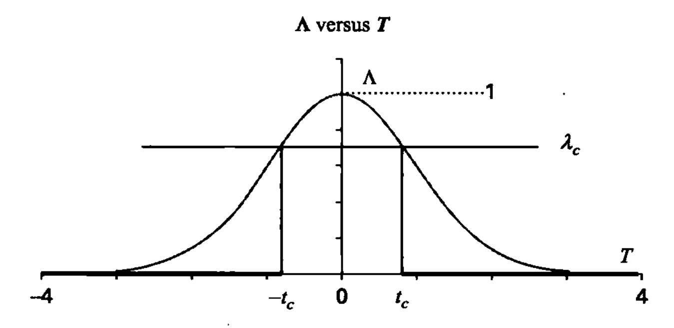


Figure 7.3-5 Instead of doing the test on the GLR statistic, it is more convenient to do the test on the T statistic. See Equation 7.3-13. The critical region along the T-axis is shown in heavy lines. For the reader's interest, for this graph $m = n =$ 10. The hypothesis is rejected if $|T| > t_c$, where t_c depends on the type I error α. In a two-sided test at significance α we assign $\alpha/2$ error mass to each half of the critical region, that is, $P[T > t_c] = \alpha/2$ and $P[T < -t_c] = \alpha/2$.

is distributed as $N(0, 1)$. Likewise,

$$W_{m+n-2} \triangleq \left(\sum_{i=1}^{m} \left(\frac{X_{1i} - \hat{\mu}_1}{\sigma}\right)^2 + \sum_{i=1}^{n} \left(\frac{X_{2i} - \hat{\mu}_2}{\sigma}\right)^2\right)$$

is Chi-square with DOF $m + n - 2$ by (iii) of Example 7.3-6. Finally, recall that $T_{m+n-2} = \frac{Z\sqrt{m+n-2}}{\sqrt{W_{m+n-2}}}$ is the t-distributed RV with DOF $m + n - 2$ so that

$$\Lambda = \left[1 + \left(T^2_{m+n-2}/(m + n - 2)\right)\right]^{-(m+n)/2}. \tag{7.3-13}$$

Since Λ is a monotonically decreasing function of T^2_{m+n-2}, the test can be made on T^2_{m+n-2} rather than on Λ. Then the critical region for H_1 of the form $0 < \Lambda < \lambda_c$ translates, when

the test is done on T^2_{m+n-2}, as the critical region (t_c^2, ∞) (Figure 7.3-4) or, equivalently, as the union of the events (t_c, ∞) and $(-\infty, -t_c)$ (Figure 7.3-5). More information on this type of test, so-called t-test, can be found in [7-10] to [7-15] and/or on the Internet by entering *t-test* in Google™ or another search engine.

Under the constraint of a type I error α we reject the hypothesis if the event $\{T^2 > t^2_{1-\alpha/2}\}$ occurs, where $t_{1-\alpha/2}$ is obtained from the t-distribution tables with $m+n-2$ degrees of freedom using $F_T(t_{1-\alpha/2}) = 1 - \alpha/2$.

Example 7.3-8

(*numerical example of testing* $H_1 : \mu_1 = \mu_2$ *versus* $H_2: \mu_1 \neq \mu_2$) We call on a Gaussian random number generator (these are available on the Internet) and generate 15 samples from a $N(0, 2)$ population (P1) and 15 samples from a $N(2, 2)$ population (P2). We reproduce the numbers here:

From population P1: $S_1 = \{2.21, 0.83, 0.393, 0.975, 0.195, -0.069, -1.91, 1.44, -3.98, 0.98, -2.84, -1.56, -0.4, -1.08, 0.116\}$; $\hat{\mu}_1' = -0.258$; $m = 15$; $\sum_{i=1}^{15} (X_{1i}' - \hat{\mu}_1')^2 = 40.48$.

From population P2: $S_2 = \{-1.28, -0.258, -0.947, 5.85, 1.56, 1.48, 1.95, 3.22, 1.41, 1.84, 2.69, 3.94, 2.04, 2.08, 1.44\}$; $\hat{\mu}_2' = 1.801$; $n = 15$; $\sum_{i=1}^{15} (X_{2i}' - \hat{\mu}_2')^2 = 45.66$. We insert the data in

$$T^2 \triangleq (m+n-2)\frac{mn(m+n)^{-1}(\hat{\mu}_1 - \hat{\mu}_2)^2}{\sum_{i=1}^{m} (X_{1i} - \hat{\mu}_1)^2 + \sum_{i=1}^{n} (X_{2i} - \hat{\mu}_2)^2} \tag{7.3-14}$$

and obtain the realization for T^2 as 10.34. Finally with $\alpha = P(\text{reject } H_1|H_1 \text{ true}) = 0.01$, we find that $F_T(t_{1-\alpha/2}) = 1 - 0.005 = 0.995$ with DOF of $15+15-2 = 28$. From the tables of the t-distribution we find that $t_{1-\alpha/2} = 2.763$ or $t^2_{1-\alpha/2} = 7.63$. Since $T^2 > t^2_{1-\alpha/2}$, we reject the hypothesis that the means are the same.

Testing for the Equality of Variances for Normal Populations: the F-test

Another problem we encounter is whether two Normal populations have the same variance. The model is the following: We have two Normal populations P1, $N(\mu_1, \sigma_1^2)$, and P2, $N(\mu_2, \sigma_2^2)$, and collect m samples (i.e., we make m i.i.d. observations) $S_1 = \{X_{1i}, i = 1, \ldots, m\}$ from P1 and n samples $S_2 = \{X_{2i}, i = 1, \ldots, n\}$ from P2. Based on these samples we wish to test the hypothesis that $H_1 : \sigma_1^2 = \sigma_2^2 \triangleq \sigma^2$ versus the alternative that $H_2 : \sigma_1^2 \neq \sigma_2^2$. The parameter space for testing H_1 is $\Theta_1 = \{\mu_1, \mu_2, \sigma^2\}$ while the parameter space for H_2 is the global parameter space $\Theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$. The likelihood function is

$$L(\Theta) = (2\pi\sigma_1^2)^{-m/2} \exp\left(-0.5\sum\nolimits_{i=1}^{m}\left(\frac{X_{1i} - \mu_1}{\sigma_1}\right)^2\right)$$

$$\times (2\pi\sigma_2^2)^{-n/2} \exp\left(-0.5\sum\nolimits_{i=1}^{n}\left(\frac{X_{2i} - \mu_2}{\sigma_2}\right)^2\right),$$

which in $\Theta_1 = \{\mu_1, \mu_2, \sigma^2\}$ assumes the form

$$L(\Theta_1) = (2\pi\sigma^2)^{-(m+n/2)} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(X_{1i}-\mu_1)^2 + \sum_{i=1}^{n}(X_{2i}-\mu_2)^2\right]\right).$$

The parameters that maximize $L(\Theta)$ in Θ_1 are, as usual, obtained by differentiating $\ln L(\Theta_1)$ with respect to μ_1, μ_2, σ^2 and setting the derivatives to zero to obtain

$$\hat{\mu}_1^* = (m)^{-1}\sum_{i=1}^{m} X_{1i} = \hat{\mu}_1; \hat{\mu}_2^* = (n)^{-1}\sum_{i=1}^{n} X_{2i} = \hat{\mu}_2;$$
$$\sigma^{2*} = (m+n)^{-1}\left(\sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^{n}(X_{2i}-\hat{\mu}_2)^2\right).$$

When these results are inserted into $L(\Theta_1)$, we obtain

$$L_{LM} = \left(\frac{2\pi}{(m+n)}\left[\sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^{n}(X_{2i}-\hat{\mu}_2)^2\right]\right)^{-(m+n)/2} \exp(-(m+n)/2).$$

To maximize $L(\Theta)$ in $\Theta = \{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$ we differentiate $\log L(\Theta)$ with respect to μ_1, μ_2, σ_1^2, σ_2^2 and set the derivatives to zero to obtain

$$\hat{\mu}_1^\dagger = (m)^{-1}\sum_{i=1}^{m} X_{1i} = \hat{\mu}_1; \hat{\mu}_2^\dagger = (n)^{-1}\sum_{i=1}^{n} X_{2i} = \hat{\mu}_2$$
$$\hat{\sigma}_1^{2\dagger} = (m)^{-1}\sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2 = \hat{\sigma}_{1,ML}^2; \hat{\sigma}_2^{2\dagger} = (n)^{-1}\sum_{i=1}^{n}(X_{2i}-\hat{\mu}_2)^2 = \hat{\sigma}_{2,ML}^2.$$

We note that the *maximum likelihood variance estimators* $\hat{\sigma}_{1,ML}^2, \hat{\sigma}_{2,ML}^2$ of the variance σ_1^2, σ_2^2 are not unbiased. When we substitute these results into $L(\Theta)$, we obtain L_{GM} as

$$L_{GM} = (2\pi)^{-(m+n)/2}\left(\frac{1}{m}\sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2\right)^{-m/2}\left(\frac{1}{n}\sum_{i=1}^{m}(X_{2i}-\hat{\mu}_2)^2\right)^{-n/2}$$
$$\times \exp(-(m+n)/2).$$

Finally, with $\Lambda = L_{LM}/L_{GM}$ we obtain

$$\Lambda = \frac{\left(\dfrac{m+n}{\left(\sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2 + \sum_{i=1}^{n}(X_{2i}-\hat{\mu}_2)^2\right)}\right)^{(m+n)/2}}{\left(\dfrac{m}{\sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2}\right)^{m/2}\left(\dfrac{n}{\sum_{i=1}^{n}(X_{2i}-\hat{\mu}_2)^2}\right)^{n/2}}. \tag{7.3-15}$$

This formidable-looking expression can be dramatically simplified by recognizing that

$$(m-1)\hat{\sigma}_1^2 = \sum_{i=1}^{m}(X_{1i}-\hat{\mu}_1)^2$$
$$(n-1)\hat{\sigma}_2^2 = \sum_{j=1}^{n}(X_{2j}-\hat{\mu}_2)^2$$



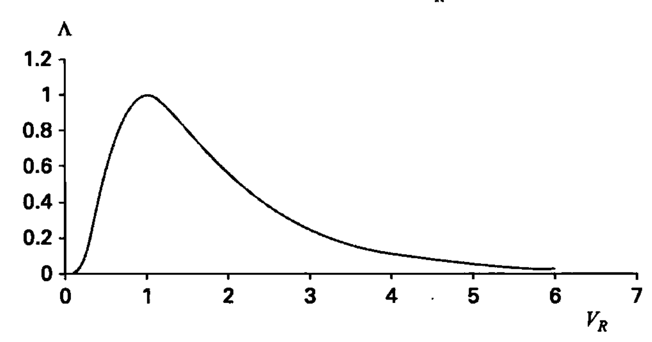


Figure 7.3-6 The test statistic Λ versus the variance ratio V_R for $m = n = 10$.

so that, after a little algebra, we obtain

$$\Lambda = A(m,n)\frac{\left([(m-1)/(n-1)] \times \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}\right)^{m/2}}{\left(1 + [(m-1)/(n-1)] \times \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}\right)^{(m+n)/2}},$$

where $A(m,n) \triangleq (m+n)^{(m+n)/2} m^{-m/2} n^{-n/2}$. It is natural to call $V_R \triangleq \hat{\sigma}_1^2/\hat{\sigma}_2^2$ the (estimator) *variance ratio*, where

$$V_R \triangleq \frac{(m-1)^{-1}\sum_{i=1}^{m}(X_i - \hat{\mu}_1)^2}{(n-1)^{-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_2)^2}. \tag{7.3-16}$$

Then, in terms of V_R,

$$\Lambda = A(m,n)\frac{([(m-1)/(n-1)] \times V_R)^{m/2}}{(1 + [(m-1)/(n-1)] \times V_R)^{(m+n)/2}} \triangleq \Lambda(V_R). \tag{7.3-17}$$

When H_1 is true $V_R = F_{m-1,n-1}$, where $F_{m-1,n-1}$ is the random variable with the F-distribution with $m-1$ and $n-1$ degrees of freedom, respectively. The variation of Λ with V_R is shown in Figure 7.3-6 for $m = n = 10$. It should be clear from the figure that rejection of the hypothesis, that is, the event $\{0 < \Lambda(V_R) < c\}$, is equivalent to the two-tailed event $\{0 < V_R < t_l\} \cup \{t_u < V_R < \infty\}$. Hence, given a significance level α, we can solve for t_l and t_u from $P[0 < V_R < t_l] + P[t_u < V_R < \infty] = \alpha$, using $\Lambda(t_l) = \Lambda(t_u)$. But for simplicity and without much loss of accuracy, we choose $P[0 < V_R < t_l'] = P[t_u' < V_R < \infty] = \alpha/2$, the numbers t_l' and t_u' being easier to determine than the numbers t_l and t_u. See Figure 7.3-7. Indeed with $F_F(x_\beta; m-1; n-1)$ denoting the CDF of the RV $F_{m-1,n-1}$ evaluated at the β percentile point, that is, $F_F(x_\beta; m-1; n-1) \triangleq \beta$, we observe that $t_l' = x_{\alpha/2}$ and $t_u' = x_{1-\alpha/2}$.

The hypothesis H_1 is rejected when the test yields the event $\{0 < \Lambda < c\}$ or, equivalently, when $\{0 < V_R < x_{\alpha/2}\}$ or $\{x_{1-\alpha/2} < V_R\}$.

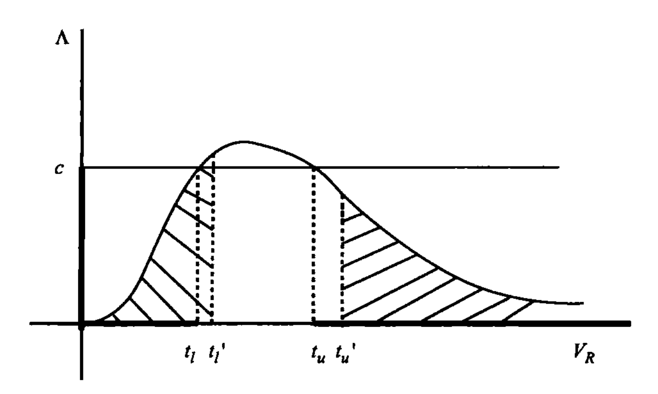


Figure 7.3-7 The event $\{0 < \Lambda < c\}$ is equivalent to the event $\{0 < V_R < t_l\} \cup \{t_u < V_R < \infty\}$. The numbers t_l and t_u are replaced by numbers t_l' and t_u' that make the error in both tails $\alpha/2$.

Example 7.3-9

(*numerical example of testing* $\mathrm{H}_1 : \sigma_1^2 = \sigma_2^2 \triangleq \sigma_0^2$ *versus* H_2: $\sigma_1^2 \neq \sigma_2^2$) We test the hypothesis that the variances of two populations are the same.

We call the RANDOM.ORG routine available on the Internet and create two sets of Gaussian pseudo-random numbers as shown in the two rows below:

N(0,1):	0.436,	−1.06,	−1.11,	0.46,	0.491,	−1.05,	0.502,	0.598,	1.61,
				−0.981,	−0.021,	0.253,	−1.24,	0.059,	2.12;
N(0,4):	0.634,	0.0818,	−1.32,	2.96,	3.11,	3.13,	2.62,	−1.96,	0.85,
				−6.51,	−3.39,	4.25,	−1.08,	3.42,	2.72;

From the top two rows, that is, the $(N(0,1))$ data we compute $\hat{\mu}_1 = 0.074, \hat{\sigma}_1 = 1.01, \hat{\sigma}_1^2 = 1.01$; from the bottom two rows, that is, the $(N(0,4))$ data we compute $\hat{\mu}_2 = 0.54, \hat{\sigma}_2 = 3.04, \hat{\sigma}_2^2 = 9.25$. We compute the variance ratio as

$$V_R = \frac{(15-1)\sum_{i=1}^{15}(X_{1i} - 0.54)^2}{(15-1)\sum_{i=1}^{15}(X_{2i} - 0.074)^2} = \frac{9.25}{1.02} = 9.06.$$

At the level of $\alpha = 0.05$, and using the "equal-area" system for distributing the error probability, we seek the percentile points numbers $x_{0.025}$ and $x_{0.975}$ such that $F_{\mathrm{F}}(x_{0.025}; 14; 14) = 0.025$ and $F_{\mathrm{F}}(x_{0.975}; 14; 14) = 0.975$. As an alternative to using F-distribution tables, we call the Stat Trek Online Statistical Table for the F-distribution calculator, and enter the degrees of freedom (14 in both cases) and the CDF value of 0.025 to obtain $x_{0.025} = 0.34$. We repeat with the CDF value of 0.975 and obtain $x_{0.975} = 2.98$. Thus, the acceptance region is the interval (event) $(0.34, 2.98)$ and the critical region is the event $\{(0,\ 0.34) \cup (2.98, \infty)\}$. The test statistic yields 9.06, an event deep in the rejection region and therefore associated with the rejection of the hypothesis that the two variances are the same. Therefore we conclude, quite rightly, that the data come from different populations.

More on the so-called F-test can be found in [7-5] to [7-9] and online by a Google search on the entry "F-test."

Testing Whether the Variance of a Normal Population Has a Predetermined Value

In this situation we consider a Normal population and test whether the variance of this population has a predetermined value. We proceed as follows: We take m samples from a Normal population X, that is, make m i.i.d. observations on X that we label: $\{X_i, i = 1, \ldots, m\}$. Under H_1 we assume that the variance of the population is the predetermined σ_0^2. The alternative hypothesis H_2 is that the variance of the population is not equal to σ_0^2 or, more precisely, that there is not enough evidence to support the validity of H_1. *As* usual we begin with the likelihood function and maximize it, respectively, in $\Theta_1 = \{\mu, \sigma_0^2\}$ and $\Theta = \{\mu, \sigma^2\}$. Thus, $L(\Theta_1) = (2\pi\sigma_0^2)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^m \left(\frac{X_i-\mu}{\sigma_0}\right)^2\right)$, which is maximized when $\hat{\mu}^* = \hat{\mu} \triangleq (m)^{-1}\sum_{i=1}^m X_i$. Thus,

$$L_{LM} = (2\pi\sigma_0^2)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^m \left(\frac{X_i-\hat{\mu}}{\sigma_0}\right)^2\right).$$

Likewise,

$$L(\Theta) = (2\pi\sigma^2)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^m \left(\frac{X_i-\mu}{\sigma}\right)^2\right),$$

which is maximized when $\hat{\mu}^\dagger = \hat{\mu} \triangleq (m)^{-1}\sum_{i=1}^m X_i$ and $\hat{\sigma}^{2\dagger} = \hat{\sigma}^2 = (m)^{-1}\sum_{i=1}^m (X_i - \hat{\mu})^2$. Hence

$$L_{GM} = (2\pi\hat{\sigma}^2)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^m \left(\frac{X_i-\hat{\mu}}{\hat{\sigma}}\right)^2\right).$$

The generalized likelihood ratio is then

$$\begin{aligned}\Lambda &= L_{LM}/L_{GM} \\ &= \left((m)^{-1}\sum_{i=1}^m \left(\frac{X_i-\hat{\mu}}{\sigma_0^2}\right)^2\right)^{m/2} \exp\left(-0.5\sum_{i=1}^m \left(\frac{X_i-\hat{\mu}}{\sigma_0^2}\right)^2 + m/2\right).\end{aligned}$$

We note that $W \triangleq \sum_{i=1}^m ((X_i - \hat{\mu})/\sigma_0)^2$ is χ^2_{m-1}. Then

$$\Lambda = \left((m)^{-1}W\right)^{m/2} \exp\left(-0.5\left(W - m\right)\right),$$

which is graphed as W versus Λ in Figure 7.3-8 for a DOF $= 9$.

From Figure 7.3-8 we deduce that the critical event, that is, the event $\{0 < \Lambda < c\}$, is equivalent to $\{0 < W < t_l\} \cup \{t_u < W < \infty\}$, where $\Lambda(t_l) = \Lambda(t_u)$ and $t_l < t_u$. For simplicity, however, we might choose the "equal area" rule by which we seek numbers $t_l' < t_u'$ such that $t_l' = x_{\alpha/2}$ and $t_u' = x_{1-\alpha/2}$, where $x_{\alpha/2}$ and $x_{1-\alpha/2}$ are α / 2 and $1-(\alpha/2)$ percentiles, that is, $F_{\chi^2}(x_{\alpha/2}; m-1) = \alpha/2$ and $F_{\chi^2}(x_{1-\alpha/2}; m-1) = 1 - (\alpha/2)$ and, as usual, $\alpha = P[\text{reject } H_1|H_1 \text{ true}]$.



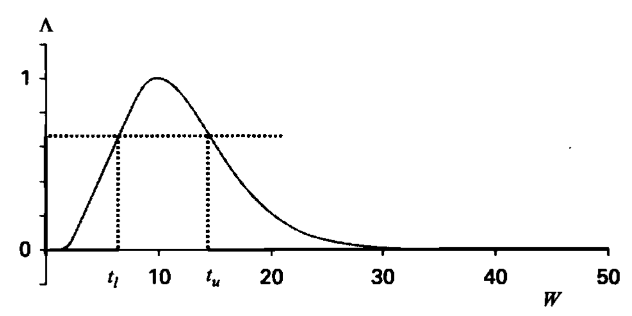


Figure 7.3-8 The critical region for Λ, shown in heavy line along the ordinate, can be related to a two-sided critical region on W (shown in heavy lines along the abscissa).

Example 7.3-10

(*numerical example of testing* $\text{H}_1 : \sigma^2 = \sigma_0^2$ *versus* H_2: $\sigma^2 \neq \sigma_0^2$) For testing purposes we draw two sets of Normal random numbers from populations we call P1 and P2, respectively. The P1 population is $N(1, 1)$ while the P2 population is $N(1, 4)$. We shall test both populations for the hypothesis that $\sigma^2 = 1$. The numbers are from RANDOM.ORG available on the Internet:

N(1,1) [P1]	−0.0644	2.91	−0.323	1.21	2.66	0.45	1.26	0.923	1.96	1.62
N(1,4) [P2]	0.705	0.685	0.718	1.03	2.52	1.96	0.417	2.69	−1.52	2.98

From the P1 data we compute $W' = 10.3$. At the 0.05 level of significance the critical region is the event $\{0 < W < 2.7\} \cup (19 < W < \infty\}$. Since W is outside this region, we accept the hypothesis that the variance of the P1 population is one. We repeat the experiment using the P2 data. Here we compute $W' = 16.5$; this is still in the acceptance region (barely) so we accept the hypothesis (in error) that the variance of the P2 population is one. We repeat the experiment at the 0.2 level of significance and find that the critical region is the event $\{0 < Z < 4.17\} \cup (14.7 < Z < \infty\}$. We find that we still accept the hypothesis that P1 has a variance of one but reject the hypothesis that P2 has a variance of one. There are two points to be made from this example: (1) Small sample sizes can lead to errors and any results drawn from them should be viewed some skepticism; (2) recalling the meaning of α, we see that if this parameter is chosen to be very small, the critical region becomes very small so that rejection of the hypothesis becomes unlikely.

7.4 GOODNESS OF FIT

An important problem in statistics is to test whether a set of probabilities have a predetermined set of specified values. For example, suppose we wish to determine whether observed data come from a standard Normal distribution. Then from the Normal tables of the function F_{SN} we can compute probabilities of the form $p_i = F_{SN}(x_{i+1}) - F_{SN}(x_i), i = 1, \ldots, l$, and compare these numbers with data obtained from multiple, independent observations on

an RV X. We can test other distributions in the same way, be they discrete or continuous. The general model is that of sorting the data into l "bins" and comparing for $i = 1, \ldots, l$ the estimated probability $\hat{p}_i$ with the specified probability p_i. Typically, if Y_i denotes the number of outcomes in n trials classified as belonging to "bin" i, then $\hat{p}_i = Y_i/n$. If all of the $\{\hat{p}_i\}$ are close to the corresponding $\{p_i\}$, it is likely that the data come from a population that has the predetermined probabilities. However, if two or more of the $\hat{p}_i$ are far from the corresponding p_i, we cannot conclude that the tested population has the same parameters as the assumed one. The choice of the number of "bins," say l, for a discrete random variable with a finite number of outcomes (the elements of the sample space) is typically the number of outcomes; thus for a die, l would be six, and for a coin, l would be two. When we deal with continuous random variables, the "bins" become intervals (x_i, x_{i+1}) $i = 1, \ldots, l$ associated with the l outcomes of the form $\{x_i < X < x_{i+1}, i = 1, \ldots, l\}$. Now the choices of l requires more thought. How "refined" a test do we need? A refined test, that is, one that contains many bins, will typically need far more data than are bins. Acquiring so much data may be costly or unrealistic. However, if we choose to make an "unrefined" test, that is, select a small number of bins, our test will necessarily be coarse. Alternatively, a large number of bins with insufficient data can lead to gross errors and make our test meaningless.

Such considerations are, more properly, in the province of experimental design and data processing. As such they are beyond the scope of the material in this book.

In the goodness-of-fit test, the hypothesis H_1 is that a set of probabilities $\{p_i, i = 1, \ldots, l\}$ satisfies $\{p_i = p_{0i}, i = 1, \ldots, l\}$. The given probabilities $\{p_{0i}, i = 1, \ldots, l\}$ characterize a probability function such as a distribution function, the outcome probabilities of a fair die, etc. We make n i.i.d. observations on an RV X and sort them into l bins depending on their values.

We define an RV X_{ij} as

$$X_{ij} \triangleq \begin{cases} 1, & \text{if the } j\text{th observation of } X \text{ is in bin } i \\ 0, & \text{else} \end{cases}.$$

We define $P[X_{ij} = 1] \triangleq p_i$ independent of j for $i = 1, \ldots, l$ because of the i.i.d. constraint. The RVs

$$Y_i \triangleq \sum_{j=1}^{n} X_{ij}, i = 1, \ldots, l$$

denote the number of outcomes in the bin $i = 1, \ldots, l$ from n trials. Note that $\sum_{i=1}^{l} Y_i = n$ and $\sum_{i=1}^{l} p_i = 1$. The reader will recognize this as the multinomial law discussed in Section 4.8, that is,

$$P[Y_1 = r_1, Y_2 = r_2, \ldots, Y_l = r_l] = P(\mathbf{r}; n, \mathbf{p}) = \frac{n!}{r_1! r_2! \cdots r_l!} p_1^{r_1} p_2^{r_2} \cdots p_l^{r_l}$$

$$\approx \frac{\exp\left(-\frac{1}{2}\left[\sum_{i=1}^{l}\left(\frac{r_i - np_i}{\sqrt{np_i}}\right)^2\right]\right)}{\sqrt{(2\pi n)^{l-1} p_1 p_2 \cdots p_l}}, \text{ when } n >> 1. \tag{7.4-1}$$

The pdf for the jth trial is $P_j = \prod_{i=1}^{l} p_i^{x_{ij}}$, where $\sum_{i=1}^{l} x_{ij} = 1$, $\sum_{i=1}^{l} p_i = 1$, and x_{ij} is restricted to 0 or 1. The likelihood function associated with n repeated trials is $L(\mathbf{p}) \triangleq L(p_1, \cdots, p_l) = \prod_{i=1}^{l} p_i^{X_{i1}} \prod_{i=1}^{l} p_i^{X_{i2}} \cdots \prod_{i=1}^{l} p_i^{X_{in}} = \prod_{i=1}^{l} p_i^{Y_i}$. Under $H_1 : p_i = p_{0i}, i = 1, \ldots, l$, the local maximum of the likelihood function, L_{LM}, is merely $L(\mathbf{p_0}) = \prod_{i=1}^{l} p_{0i}^{X_{i1}} \prod_{i=1}^{l} p_{0i}^{X_{i2}} \cdots \prod_{i=1}^{l} p_{0i}^{X_{il}} = \prod_{i=1}^{l} p_{0i}^{Y_i}$. The global maximum of the likelihood function is obtained by differentiation with respect to the $p_i, i = 1, \ldots, l$, while recalling that $\sum_{i=1}^{l} p_i = 1$. The result is $\hat{p}_i = Y_i/n, i = 1, \ldots, l$. Thus, $L_{GM} = L(Y_1/n, Y_2/n, \cdots, Y_l/n) = \prod_{i=1}^{l} (Y_i/n)^{Y_i}$. Finally, recalling that $\sum_{i=1}^{l} Y_i = n$, we find that the generalized likelihood ratio is

$$\Lambda = n^n \prod_{i=1}^{l} \left(\frac{p_{0i}}{Y_i}\right)^{Y_i} \tag{7.4-2}$$

and the critical region is $0 < \Lambda < \lambda_c$. To compute the critical region at a specified level of significance, we need the distribution of Λ. However, the exact distribution of Λ under H_1 for an arbitrary value of n is difficult to obtain. It is shown elsewhere [7-1] that $-2 \ln \Lambda$ under the large sample assumption is approximately χ^2_{l-1}.

We consider here another approach. From Equation 7.4-1 we see that the $Y_i, i = 1, \ldots l$, under the large sample assumption can be approximated by Normal RVs $N(np_i, np_i), i = 1, \ldots l$, while the $U_i \triangleq \frac{Y_i - np_i}{\sqrt{np_i}}, i = 1, \ldots l$, are approximately standard Normal. Now consider the test statistic

$$V \triangleq \sum_{i=1}^{l} \left(\frac{Y_i - np_{0i}}{\sqrt{np_{0i}}}\right)^2, \tag{7.4-3}$$

which is called the *Pearson test statistic*, and accepting or rejecting a hypothesis based on the size of V is called *Pearson's test* or the Chi-square test [7-16] to [7-20]. Pearson's test statistic has the form of a χ^2 RV with l degrees of freedom but, in fact, has only $l-1$ degrees of freedom because $Y_l = n - \sum_{i=1}^{l-1} Y_i$ is completely specified once the $Y_1, Y_2, \ldots, Y_{l-1}$ are specified. Now, if the Y_i come from a population with probabilities $p_{0i}, i = 1, \ldots, l$, we expect that a realization of V will be small. However, if the Y_i come from a population with probabilities $p_i, i = 1, \ldots, l$, where at least two of the p_i are significantly different from the corresponding p_{0i}, we expect realizations of V to be large. We can demonstrate this by computing $E[V]$ under H_1 and H_2. Under H_1 we compute $E[V|H_1] = l - 1$ (see Problem 7.24). However, under H_2 we compute $E[V|H_2]$ as

$$E[V|H_2] \approx \sum_{i=1}^{l} (p_{0i})^{-1} n (p_{1i} - p_{0i})^2 \tag{7.4-4}$$

when n is large (see Problem 7.25). Clearly $E[V|H_2]$ can become arbitrarily larger than $l - 1$ when at least some of the p_{1i} are different from p_{01}. An exact computation of $E[V|H_2]$ would show that it can never be smaller than $l - 1$.

Returning to the test statistic in Equation 7.4-3, that is,

$$V \triangleq \sum_{i=1}^{l} \left(\frac{Y_i - np_{0i}}{\sqrt{np_{0i}}} \right)^2$$

we note that under H_1 it is χ^2_{l-1}. To find the constant c that determines the critical region $\{V > c\}$ at significance α, we solve $\int_c^\infty f_{\chi^2}(x; l-1)dx = \alpha$ or, equivalently, $1 - \alpha = F_{\chi^2}(c; l-1)$. So we find that $c = x_{1-\alpha}$, the $1-\alpha$ percentile point of χ^2_{l-1}. Thus our criterion becomes: accept H_1 if $V < x_{1-\alpha}$, else reject H_1.

Example 7.4-1

(*fairness of a coin*) We wish to test the hypothesis H_1 that $p_{01} = P[heads] = 0.5 = p_{02} = P[tails]$ at a level of significance $\alpha = 0.05$. We flip the coin 100 times and observe 61 heads and 39 tails. Then from

$$V \triangleq \sum_{i=1}^{l} \left(\frac{Y_i - np_{0i}}{\sqrt{np_{0i}}} \right)^2,$$

we obtain $V' = \frac{1}{0.5 \times 100}[61 - 50]^2 + \frac{1}{0.5 \times 100}[39 - 50]^2 = 4.84$. We compute the critical value from $0.95 = F_{\chi^2}(x_{0.95}; 1)$, which yields $x_{0.95} = 3.84$. Since $V' = 4.84 > 3.84$ *we* reject the hypothesis that the coin is fair.

Example 7.4-2

(*fairness of a die*) We wish to test the hypothesis, at significance 0.05, that a six-faced die is fair. We let $Y_i, i = 1, \ldots, 6$, denote the number of times face i shows up. We cast the die 1000 times and observe $Y_1' = 152, Y_2' = 175, Y_3' = 165, Y_4' = 180, Y' = 159, Y_6' = 171$. Then

$$V' = \frac{1}{167}\Big[(167 - 152)^2 + (167 - 175)^2 + (167 - 165)^2 + (167 - 180)^2 + (167 - 159)^2 + (167 - 171)^2\Big] = 3.25.$$

The degree of freedom is five so we solve $0.95 = F_{\chi^2_5}(x_{0.95})$. This yields $x_{0.95} = 11.1$ and since $3.25 < 11.1$ we accept the hypothesis that the die is fair.

Example 7.4-3

(*test of Normality*) We wish to determine whether data are from a standard Normal $N(0,1)$ population. We let H_1 be the hypothesis that X is a distributed as a standard Normal $N(0,1)$ and H_2 be the alternative that X is not distributed as $N(0,1)$. We use differences of the cumulative Normal distribution for the $\{p_{0i}\}$ as follows:

$p_{01} \triangleq F_{SN}(-2.0) = 0.023; p_{02} \triangleq F_{SN}(-1.5) - F_{SN}(-2.0) = 0.044; p_{03} \triangleq F_{SN}(-1.0) - F_{SN}(-1.5) = 0.092; p_{04} \triangleq F_{SN}(-0.5) - F_{SN}(-1.0) = 0.145; p_{05} \triangleq F_{SN}(0) - F_{SN}(-0.5) = 0.1915; p_{06} \triangleq F_{SN}(0.5) - F_{SN}(0) = 0.1915; p_{07} \triangleq F_{SN}(1.0) - F_{SN}(0.5) = 0.15; p_{08} \triangleq F_{SN}(1.5) - F_{SN}(1.0) = 0.092; p_{09} \triangleq F_{SN}(2.0) - F_{SN}(1.5) = 0.044; p_{010} \triangleq F_{SN}(\infty) - F_{SN}(2) = 0.023.$

In a 1000 observations we observe the following realizations:

in the interval$(-\infty, -2]$: $Y_1' = 19$
in the interval$(-2, -1.5]$: $Y_2' = 42$
in the interval$(-1.5, -1]$: $Y_3' = 96$
in the interval$(-1, -0.5]$: $Y_4' = 135$
in the interval$(-0.5, 0]$: $Y_5' = 202$
in the interval$(0,\ 0.5]$: $Y_6' = 193$
in the interval$(0.5,\ 1]$: $Y_7' = 155$
in the interval$(1, 1.5]$: $Y_8' = 72$
in the interval$(1.5, 2]$: $Y_9' = 53$
in the interval$(2, \infty]$: $Y_{10}' = 33$

We use $V \triangleq \sum_{i=1}^{10} \left(\frac{Y_i - 1000p_{0i}}{\sqrt{1000p_{0i}}} \right)^2$ as the test statistic and observe that V is χ_9^2 if H_1 is true. From the given data compute $V' = 12.9$. Since $x_{0.95} = 16.92$ and 12.9 is less than 16.92 we accept the hypothesis that the data are Normally distributed.

We can use Pearson's test statistic to test whether two *unknown* probabilities are equal even if no other prior information such as means and variances is available. For example we test two brands of printing paper in printing presses: Brand A clogs the presses six times in 150 trials while brand B clogs the presses 25 times in 550 trials. Are brands A and B equally likely to clog the presses? Two speech recognition programs are available for purchase. Assuming the same speaker, we find that speech recognition program SR1 mistakes 61 words out of 250 while SR2 mistakes 30 words out of 110. Are both programs equally effective? In the framework of probability theory we model this as follows: We consider the occurrences of two events say E_1 and E_2 and we ask, Is $P[E_1] = P[E_2]$? Define Z_1 as the number of times we observe the occurrence of E_1 in m trials and Z_2 as the number of times we observe the occurrence of E_2 in n subsequent trials. Let $p_1 \triangleq P[E_1]$ and $p_2 \triangleq P[E_2]$. Let $m >> 1$, $n >> 1$, then by the Central Limit Theorem $Z_1{:}N(mp_1, mp_1q_1)$ and $Z_2{:}N(np_2, np_2q_2)$. We define the normalized RVs $Y_1 \triangleq Z_1/m{:}N(p_1, p_1q_1/m)$, and $Y_2 \triangleq Z_2/n{:}N(p_2, p_1q_1/n)$ and consider the RV $Y \triangleq Y_1 - Y_2$. Since $Y_1$and Y_2 are independent (recall that Y_1 results from observations in the first m trials while Y_2 results from observations in the next n trials), it follows that Y is Normal with mean $p_1 - p_2$ and variance $\sigma_Y^2 = (np_1q_1 + mp_2q_2)/mn$. Let H_1 be the hypothesis that $p_1 = p_2$ and the alternative H_2 be that $p_1 \neq p_2$; clearly under H_1, $Y{:}N\,(0, p_1q_1(m+n)/nm)$. The Pearson test statistic adapted to this problem is

$$V = \left(\frac{Y - (p_1 - p_2)}{\sigma_Y} \right)^2,$$

which is seen to be χ_1^2. For a test of significance α we find the percentile $x_{1-\alpha}$ in $1 - \alpha = F_{\chi^2}(x_{1-\alpha}; 1)$ such that if $V < x_{1-\alpha}$ we accept the hypothesis; else we reject the hypothesis.

The difficulty with this problem is that σ_Y is unknown since p_1 and p_2 are unknown. One way out of this difficulty is to replace σ_Y by an estimate of σ_Y based on our observations. Under $H_1, p_1 = p_2 \triangleq p$, and the minimum variance, unbiased estimator of p is $\hat{p} = (Z_1 + Z_2)/(m+n)$. It follows that under $H_1, \hat{\sigma}_Y = \sqrt{\hat{p}\hat{q}(m+n)/mn}$, where $\hat{q} = 1-\hat{p}$. We illustrate with two examples.

Example 7.4-4

(*voting patterns in different regions*) In the Governor's race in a large state, exit polls showed that in a rural upstate county 167 out of 211 voters voted for the Republican while in a downstate county that includes a large metropolitan area, 216 out of 499 voters voted Republican. Can we assume that the probability, p_1, that an upstate voter will vote Republican is the same as, p_2, that of a downstate voter?

Solution Under $H_1, p_1 = p_2 \triangleq p$, while under $H_2, p_1 \neq p_2$. Under H_1, we compute $\hat{p}' = 388/710 = 0.54$, $\hat{q}' = 0.46$, $\hat{\sigma}'_Y = 0.041$, $Y'_1 = 167/211 = 0.79$, $Y'_2 = 216/499 = 0.43$, and $Y' \triangleq Y'_1 - Y'_2$ =0.36; hence $V' = (0.36/0.041)^2 \approx 77$. At a significance level of $\alpha = 0.05$, we find that $x_{0.95} = 3.84$. Since $77 > 3.84$, the hypothesis is strongly rejected.

Example 7.4-5

(*interpretation of scientific data*) In an attempt to find out whether Rhesus monkeys can be made to distinguish and possibly attach meaning to different sounds, including spoken language, the following experiment was performed. A Rhesus monkey was put in an anechoic (external-soundproof) chamber with a computer-controlled directional loudspeaker that randomly emitted bursts of one of two signals: S1, a sound of the type that the Rhesus monkey might hear in its natural habitat; and S2, a sound characteristic of a spoken word. If the monkey, upon hearing a sound burst, turned its head toward the loudspeaker, it was taken to mean that the monkey was reacting to the sound. If the sound was of an S2 type, it could mean that the monkey was curious or interested in the sound and could possibly be trained to accept the sound as a word. However, if the monkey showed no reaction to the sound, it was taken to mean that the monkey attached no significance to it. From the researcher's point of view the ideal case would be if the monkey *never* turned its head when exposed to an S1 sound and *always* turned its head when exposed to an S2 sound. Then the researcher could write a scholarly paper on the cognitive abilities of the Rhesus monkey and become famous.† We shall ignore the perplexing problem of deciding whether the monkey's head has rotated enough to be scored as a "turned head."‡

In 267 bursts of "natural habitat"-type sounds, the monkey turned its head 112 times; in 289 bursts of spoken word sounds, the Rhesus monkey turned its head 173 times. Let p_1 denote the probability that a monkey will turn its head upon hearing a "natural habitat" sound and p_2 denote the probability that the monkey will turn its head upon spoken-worn

†This research is being done at a major university but the results have generated controversy in the scientific community.

‡A problem similar to the "checked swing" problem in baseball, where the umpire must decide whether a batter "followed through" or "checked his swing."

sounds. Under $H_1, p_1 = p_2 \triangleq p$ while under $H_2, p_1 \neq p_2$. Can we accept the hypothesis that the monkey shows no differentiation in its reaction to the two sounds, that is, that $H_1, p_1 = p_2 \triangleq p$, is true?

Solution Under $H_1, \hat{p}' = 0.51$, $\hat{q}' = 0.49, \hat{\sigma}' = 0.0424$, $Y_1' = 112/267 = 0.42$, $Y_2' = 173/289 = 0.6$, and $Y' \triangleq Y_1' - Y_2' = 0.18$; hence $V' = (0.18/0.0424)^2 = 18$; at the 0.05 level of significance $x_{0.95} = 3.84$. Hence the hypothesis is strongly rejected.

7.5 ORDERING, PERCENTILES, AND RANK

For the reader's convenience we repeat here some of the material from Section 6.8 of chapter 6. We make n i.i.d. observations on a generic RV X (sometimes called a *population*) with CDF $F_X(x)$ to obtain the sample $X_1, X_2, \ldots, X_n$. The joint pdf of the sample is $f_X(x_1) \times \cdots \times f_X(x_n)$, $-\infty < x_i < \infty, i = 1, \ldots, n$. Next we order the $X_i, i = 1, \ldots, n$, by size (signed magnitude) to obtain the *ordered sample* $Y_1, Y_2, \ldots, Y_n$ such that $-\infty < Y_1 < Y_2 < \cdots < Y_n < \infty$. This is sometimes called the *order statistics* of the observations on X. When ordered, the sequence $3, -2, -9, 4$ would become $-9, -2, 3, 4$. If a sequence $X_1, \ldots, X_{20}$ was generated from n observations on $X : N(0,1)$, it would be very unlikely that $Y_1 > 0$ because this would require that the other 19 $Y_i, i = 2, \ldots, 20$, be greater than zero and therefore all the samples would be on the positive side of the Normal curve. The probability of this event is $(1/2)^{20}$. Likewise it would be extremely unlikely that $Y_{20} < 0$ because this would require that the other 19 $Y_i, i = 1, \ldots, 19$, be less than zero. As shown in Section 5.3, the joint pdf of the ordered sample $Y_1, Y_2, \ldots, Y_n$ is $n! f_X(y_1) \times \cdots \times f_X(y_n)$, $-\infty < y_1 < y_2 < \cdots < y_n < \infty$, and zero else. *Ordering* and *ranking* are not the same in that ranking normally assigns a value to the ordered elements. For example most people would order the pain of a broken bone higher than that of a sore throat due to a cold. But if a physician asked the patient to *rank* these pains on a scale of 0 to 10, the pain associated with the broken bone might be ranked at 8 or 9 while the sore throat might be given a rank of 3 or 4.

Consider next the idea of *percentiles*. We have used the notion of percentiles in other places in the book; here we briefly discuss it in greater detail. Assume that the IQ of a large segment of a select population is distributed as $N(100, 100)$, that is, a mean of 100 and a standard deviation of 10. Obviously the Normal approximation is valid only over a limited range because no one has an IQ of 1000 or an IQ of -10. The IQ test itself is valid only over a limited range and may not give an accurate score for people that are extremely bright or severely cognitively handicapped. It is sometimes said that people in either group are "off the IQ scale." Still the IQ test is widely used as an indicator of problem-solving ability. Suppose that the result of an IQ test says that the child ranks in the 93rd percentile of the examinees and therefore qualifies for admission to programs for the "gifted." How do we locate the 93rd percentile on the IQ scale?

Definition (percentile): Given an RV X with CDF $F_X(x)$, the u-percentile of X is the number x_u such that $F_X(x_u) = u$. If the CDF F_X is everywhere continuous with

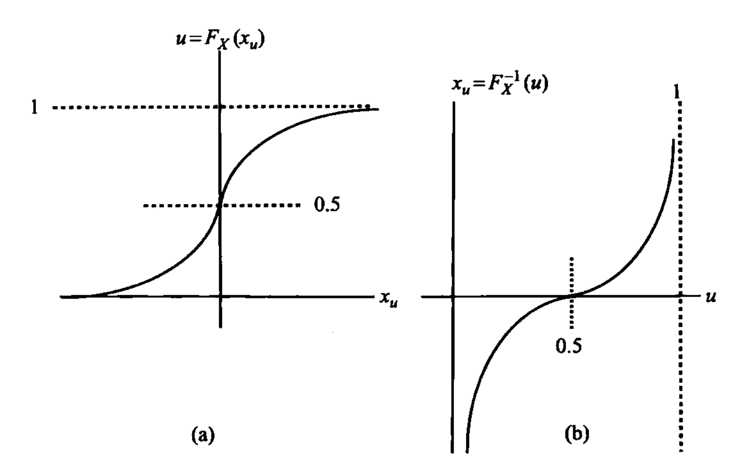


Figure 7.5-1 (a) The standard Normal CDF; (b) the inverse function.

continuous derivative, then $x_u = F_X^{-1}(u)$, where the function F_X^{-1} is the inverse function associated with the CDF F_X, that is, $F_X^{-1}(F_X(x_u)) = x_u$. The standard Normal CDF and its inverse are shown in Figure 7.5-1.

Observation In the special case of the standard Normal, where $Z : N(0,1)$, we use the symbol z_u to denote the u-percentile of X. If $X{:}N(\mu, \sigma^2)$, then the u-percentile of X, x_u, is related to z_u according to

$$x_u = \mu + z_u\sigma. \tag{7.5-1}$$

Example 7.5-1

(*relation between* $\mathrm{x_u}$ *and* $\mathrm{z_u}$) Show that $x_u = \mu + z_u\sigma$.

Solution We write

$$\begin{aligned} F_X(x_u) = u &= \left(2\pi\sigma^2\right)^{-1/2} \int_{-\infty}^{x_u} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right) dx \\ &= (2\pi)^{-1/2} \cdot \int_{-\infty}^{(x_u-\mu)/\sigma} \exp\left(-\frac{1}{2}z^2\right) dz \\ &\triangleq (2\pi)^{-1/2} \int_{-\infty}^{z_u} \exp\left(-\frac{1}{2}z^2\right) dz. \end{aligned}$$

The last line is the CDF of $Z{:}N(0,1)$. Hence $x_u = \mu + z_u\sigma$. We can use this result in the previously mentioned IQ problem. From the data we have $F_X(x_u) = 0.93 = F_Z(z_u)$. From the table of F_{SN}, we get that $z_u \approx 1.48$. Then with $x_u = \mu + z_u\sigma = 100 + 1.48\ (10)$, we get that a 93 percentile in the IQ distribution corresponds to an IQ of 115.

How Ordering is Useful in Estimating Percentiles and the Median*

We briefly review here some of the material of Section 6.8 that is associated with percentiles and the median.

The *median* of the population X is the point $x_{0.5}$ such that $F_X(x_{0.5}) = 0.5$. This is to be contrasted with the *mean* of X, written as μ_X, and defined as $\mu_X = \int_{-\infty}^{\infty} x f_X(x)dx$. The median and mean do not necessarily coincide. For example, in the case of $f_X(x) = \lambda e^{-\lambda x}u(x)$ we find that $\mu_X = 1/\lambda$ but $x_{0.5} = 0.69/\lambda$. To compute the mean of X we need $f_X(x)$, which is often not known. The mean may seem like a rather abstract parameter while the median is merely the point that divides the population in half, that is, half the population is at or below the median and half above†. The situation where $f_X(x)$ is assumed to exist and for which we can extract or estimate parameters is called the *parametric case.* Typically, in the parametric case, we might assume a form for the population density, for example, the Normal, and wish to estimate some unknown parameter of the distribution, for example, the mean μ_X. Then given n i.i.d. observations $X_1, X_2, \ldots, X_n$ on X, we *estimate* μ_X with $\hat{\mu}_X = n^{-1}\sum_{i=1}^{n} X_i$, which happens to be an unbiased and consistent estimator for the mean of many populations. Indeed it is the simple form of the mean estimator function $\hat{\mu}_X$ and the fact that if σ_X^2 is finite then $\hat{\mu}_X \to \mu_X$ for large n (see the *law of large numbers*) that make the mean so useful in many applications. The estimation of parameters in known or assumed distributions and other operations, for example, hypothesis testing involving known or assumed distributions, is known as *parametric statistics.*

The estimation of the properties and parameters of a population without any assumptions on the form or knowledge of the population distribution is known as *distribution-free, robust,* or *nonparametric statistics.* Statistics based on observations only without assuming underlying distributions are *robust* in the sense that the theorems and conclusions drawn from the observations do not change with the form of the underlying distributions. Whereas the mean and standard deviation are useful in characterizing the center and dispersion of a population in the parametric case, the *median* and *range* play this role in the nonparametric case. To estimate the median from $X_1, X_2, \ldots, X_n$, we use the order statistics and estimate $x_{0.5}$ with the *sample median* estimator

$$\begin{aligned} \hat{Y}_{0.5} &= Y_{k+1} \text{ if } n \text{ is odd, that is, } n = 2k+1 \\ &= 0.5(Y_k + Y_{k+1}) \text{ if } n \text{ is even, that is, } n = 2k. \end{aligned} \tag{7.5-2}$$

The sample median is not an unbiased estimator for $x_{0.5}$ but becomes nearly so when n is large. The dispersion in the nonparametric case is measured from the 50 *percent percentile range,* that is, $\Delta x_{0.50} \triangleq x_{0.75} - x_{0.25}$, or the 90 *percent percentile range,* that is, $\Delta x_{0.90} \triangleq x_{0.95} - x_{0.05}$, or some other appropriate range.

* Readers familiar with the contents of Section 6.8 can skip this subsection.

†Thus it is not wholly accurate to say that "half the population is below and half above" the median. Moreover the reader should be aware that the median of a sample is typically not the same as the median of the whole population.

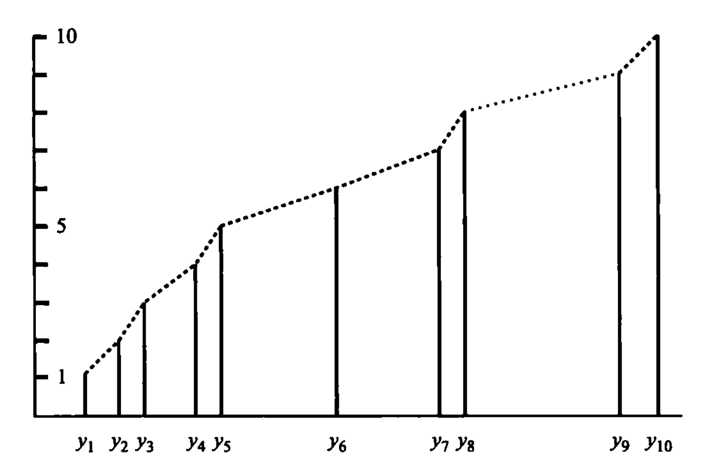


Figure 7.5-2 Estimated percentile range from ten ordered samples showing linear interpolation between the samples. To get the estimated percentile take the ordinate value and multiply by 100/11. Thus, to a first approximation, the 90th percentile is estimated from y_{10} while the 9th percentile is estimated from y_1. An approximate 50 percent range is covered by $y_8 - y_2$.

Example 7.5-2

(*interpolation to get percentile points*) Using the symbol $\alpha \sim \beta$ to mean α estimates β, we have $Y_3 \sim x_{0.273}, Y_4 \sim x_{0.364}$, and using linear interpolation, we get $x_{0.3}$ as

$$Y_4 + \frac{(Y_4 - Y_3)(0.3 - 4/11)}{1/11} \sim x_{0.3}.$$

Linear interpolation between ordered samples is illustrated in Figure 7.5-2.

We discuss next a fundamental result connecting order statistics with percentiles. Once again the model is that of collecting a sample of n i.i.d. observations $X_1, X_2, \ldots, X_n$ on an RV X with CDF $F_X(x)$. We recall the notation $P[X_i \leq x_u] \triangleq u$. Next we consider the order statistics $Y_1 < Y_2 < \cdots < Y_n$. Now consider the event $\{Y_k < x_u\}$. Since Y_k is the kth element in the ordering of the $\{X_i\}$, there are *at least* k of the $\{X_i\}$ that are less than x_u. There may be more but certainly not less. Then, because the $\{X_i\}$ are i.i.d. we can use the binomial probability formula to compute

$$P[Y_k < x_u] = P\left[\text{at least } k \text{ of the } \{X_i\} \text{ are less than } x_u\right]$$

$$= \sum_{i=k}^{n} \binom{n}{i} u^i (1-u)^{n-i}. \tag{7.5-3}$$

Next consider the event $\{Y_{k+r} > x_u\}$. Since Y_{k+r} is the $(k+r)$th element in the ordering of the $\{X_i\}$, there are *at least* $n-(k+r)+1$ of the $\{X_i\}$ that are greater than x_u. Equivalently, there can be *no more* than $k+r-1$ of the $\{X_i\}$ less than x_u. Then

$$P[Y_{k+r} > x_u] = P[\text{no more than } k+r-1 \text{ of the } \{X_i\} \text{ are less than } x_u]$$
$$= \sum_{i=0}^{k+r-1} \binom{n}{i} u^i (1-u)^{n-i}. \tag{7.5-4}$$

The intersection of the events $\{Y_{k+r} > x_u\} \cup \{Y_k < x_u\}$ is the event $\{Y_k < x_u < Y_{k+r}\}$. Its probability is

$$P[Y_k < x_u < Y_{k+r}] = \sum_{i=k}^{k+r-1} \binom{n}{i} u^i (1-u)^{n-i} \tag{7.5-5}$$

and is independent of $f_X(x)$. The result given in Equation 7.5-5 is one of the major results of nonparametric statistics and has important applications, for example, estimating the median of a population, as we illustrate below.

Example 7.5-3

(*How large a sample do we need to cover the median at 95 percent confidence?*) We seek the end points Y_1, Y_n of a random interval $[Y_1, Y_n]$ so that the event $\{Y_1 < x_{0.5} < Y_n\}$ occurs with probability 0.95. Here $Y_1 \triangleq \min(X_1, X_2, \ldots X_n), Y_n \triangleq \max(X_1, X_2, \ldots X_n)$. In effect, how large should n be?

Solution We compute

$$P[Y_1 < x_{0.5} < Y_n] = \sum_{i=1}^{n-1} \binom{n}{i} (1/2)^n \approx 0.95$$

and find that for $n = 5$, $P[Y_1 < x_{0.5} < Y_5] \approx 0.94$. The probability that the random interval $[Y_1, Y_n]$ covers the 50 percent percentile point is shown in Figure 7.5-3 for various values of n.

Example 7.5-4

(*most probable adjacent ordered pair to cover* $x_{0.33}$) We have the order statistics $\{Y_1, Y_2, \ldots, Y_n\}$ and wish to find the pair $\{Y_i, Y_{i+1}, i = 1, \ldots, n-1\}$ that maximizes the probability of covering the 33.33rd percentile point. The 33.33rd percentile point $x_{0.33}$ is defined by $1/3 = F_X(x_{0.33})$. For specificity we assume $n = 10$. From Equation 7.5-5 we compute

$$P[Y_k < x_{0.33} < Y_{k+1}] = \frac{10!}{k!(10-k)!}(1/3)^k(2/3)^{10-k}, k = 1, \ldots, 9$$

and plot the result in Figure 7.5-4. Clearly the interval $[Y_3, Y_4]$ is most likely to cover $x_{0.33}$. The probability of the event $\{Y_3 < x_{0.33} < Y_4\}$ is 0.26.

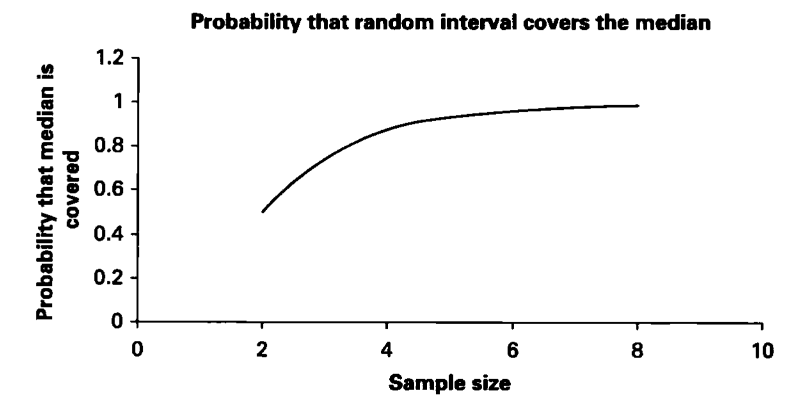


Figure 7.5-3 Probability that the event $\{Y_1 < x_{0.5} < Y_n\}$ covers the median for various values of n.

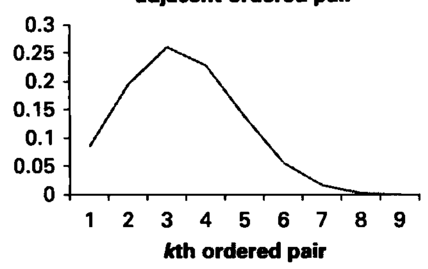


Figure 7.5-4 Among the pairwise intervals $[Y_k, Y_{k+1}]$, the interval $[Y_3, Y_4]$ is most likely to cover $x_{0.33}$. Here $n = 10$.

Example 7.5-5

(*the median and mean are not the same for the binomial*) We make the somewhat trivial observation that for the binomial case the mean and median do not coincide. For example with $p = 1/2$ and $n = 4$, the mean is 2 but the median, such as it is, is somewhere between 1 and 2. However, when n is large the median and mean approach each other and the median can be estimated by the mean. Indeed it can be shown that the error between the mean and median is proportional to $(p(1-p))^n$, which becomes arbitrarily small for $n \to \infty$.

Confidence Interval for the Median When n Is Large

If n is large enough so that the Normal approximation to the binomial is valid in distribution, we can use

$$P[\alpha \leq S_n \leq \beta] \approx \frac{1}{\sqrt{2\pi}} \int_{\alpha_n}^{\beta_n} \exp\left[-\frac{1}{2}y^2\right] dy,$$

where

$$\begin{aligned} P[\alpha \leq S_n \leq \beta] &= \sum_{i=\alpha}^{\beta} \binom{n}{i} p^i (1-p)^{n-i}, \\ \alpha_n &\triangleq \frac{\alpha - np - 0.5}{\sqrt{np(1-p)}}, \text{ and} \\ \beta_n &\triangleq \frac{\beta - np + 0.5}{\sqrt{np(1-p)}}. \end{aligned} \tag{7.5-6}$$

To apply these results to the problem at hand we write

$$P[Y_r < x_{0.5} < Y_{n-r+1}] = \sum_{i=r}^{n-r} \binom{n}{i} (1/2)^n, \tag{7.5-7}$$

where we used that, by definition of the median, $u = F_X(x_{0.5}) = 1/2$. The choice of subscripts will ensure that the confidence interval will begin at the rth place counting from the bottom, that is, 1, 2, 3,..., r, and end at the place reached by counting r observations

back from the top. For example if the 95 percent confidence calculation for $n = 10$ yields $r = 3$, the confidence interval begins at the third observation and ends at the eighth observation, both points reached by counting three places from bottom and top, respectively, that is, 1, 2, 3 (Y_3) and 10, 9, 8 (Y_8), and the result would appear as $P[Y_3 < x_{0.5} < Y_8] = 0.95$.

In the binomial sum in Equation 7.5-7 we note that its mean is $n/2$ and its standard deviation is $\sqrt{n}/2$. Hence the Normal approximation to the binomial sum in Equation 7.5-7 for a 95 confidence interval is

$$\sum_{i=r}^{n-r} \binom{n}{i} (1/2)^n \approx \frac{1}{\sqrt{2\pi}} \int_{\alpha_n}^{\beta_n} \exp[-\frac{1}{2}x^2]dx = 0.95,$$

which, from the tables of the standard Normal distribution function $F_{SN}(x)$, yields $\alpha_n = -1.96, \beta_n = 1.96$. Then it follows from Equation 7.5-6 that

$$1.96 = \frac{n - r - n/2 + 0.5}{\sqrt{n}/2}$$
$$-1.96 = \frac{r - n/2 - 0.5}{\sqrt{n}/2},$$

which yields $r = (n/2) - 1.96\sqrt{n}/2 + 0.5$. If r is not an integer replace r by $\lfloor r \rfloor$, which is the least integer function, that is, that largest integer less than or equal to r.

Example 7.5-6

(*95 percent confidence interval for the median for* n $= 20$) We make 20 observations on an RV X and label these $\{X_i, i = 1, \ldots, 20\}$. We order them by size so that $Y_1 < Y_2 < \cdots < Y_n$. We use $r = (n/2) - 1.96\sqrt{n}/2 + 0.5$ to obtain $r = 6.12$ and $\lfloor r \rfloor = 6$. Then $P[Y_6 < x_{0.5} < Y_{15}] \geq 0.95$.

Distribution-Free Hypothesis Testing: Testing If Two Populations Are the Same Using *Runs*

In general, hypothesis testing using nonparametric statistics is more involved than in the parametric case because of the difficulty of computing the distribution of the test statistic. However, when the size of the samples is large, say greater than 10, we can use the Normal approximation for computing the acceptance/rejection region.

We introduce the idea of a *run* by considering the following simple situation. We make n_1 observations on an RV X (the "population") with CDF $F_X(x)$ and label these samples $\{X_i^{(1)}, i = 1, \ldots, n_1\}$. After ordering them by size we create the samples $\{Y_i, i = 1, \ldots, n_1\}$. Then we make n_2 observations on the *same* RV X and label these samples $\{X_i^{(2)}, i = 1, \ldots, n_2\}$. We order these samples by size to obtain the ordered set $\{Z_i, i = 1, \ldots, n_2\}$. Next we combine the two unordered sets of samples into a single set and order them by size. Then a typical ordered sequence might be $Z_1, Z_2, Y_1, Z_3, Y_2, \ldots, Z_{n_2}, Y_{n_1-1}, Y_{n_1}$, where $Z_1 < Z_2 < Y_1 < Z_3 < Y_2 < \cdots < Z_{n_2} < Y_{n_1-1} < Y_{n_1}$. We define a *run* as a sequence of letters of the same kind bounded by letters of the other kind or the beginning/end of the entire sequence. Thus Z_1, Z_2 is the first run and its length is two. The next run is Y_1 and it has length one,

etc. The last run is $Y_{n_1-1}Y_{n_1}$ and it has length two. We count the total number of runs and call this D. We note that D is a random variable. Since the two sets of samples come from the *same population*, we expect a thorough mixing of the Y's and Z's and therefore a large D. Note, however, that had the Y's and Z's come from *different populations*, D would, in all likelihood, be significantly reduced. For, example, suppose that we have two populations, say, $X^{(1)}$ with pdf $f_{X^{(1)}}(x) = \text{rect}(x)$ and $X^{(2)}$ with pdf $f_{X^{(2)}}(x) = \text{rect}(x-2)$. If $\{Y_i, i = 1, \ldots, n\}$ represent the ordered sequence from the $X^{(1)}$ population and $\{Z_i, i = 1, \ldots, n\}$ represent the ordered sequence from the $X^{(2)}$ population, then the ordered samples of the mixed sequence will appear as $Y_1Y_2 \cdots Y_nZ_1Z_2 \cdots Z_n$ and will have $D' = 2$ since the support of their pdf's don't overlap.

Example 7.5-7

(*realizations of D for populations of equal and different means*) We generate two sets of ten Normal random numbers (we show only to two places) from $N(0,1)$ obtained from RANDOM.ORG, a Normal random number provider available on the Internet.

$$N(0,1) \to \{x^{(1)}: -0.19,\ 0.99,\ -1.1,\ -1.0,\ -1.3,\ -0.53,\ -0.25,\ 0.75,\ -0.25,\ 0.75\}$$
$$N(0,1) \to \{x^{(2)}: 0.68,\ -1.2,\ 0.28,\ 0.61,\ -1.2,\ -1.5,\ 2.1,\ -0.10,\ -0.87,\ 0.80\}.$$

We order by size the $x^{(1)}$ and $x^{(2)}$ sequences separately to create, respectively, the ordered sequences $y_1y_2 \cdots y_{10}$ and $z_1z_2 \cdots z_{10}$, where $y_1 = -1.3$, $y_{10} = 0.99$, $z_1 = -1.5$, and $z_{10} = 2.1$. After combining the two sequences into a single sequence and ordering all the elements of this sequence by size, we get the sequence $z_1y_1z_2z_3y_2y_3z_4y_4y_5y_6y_7z_5z_6z_7z_8y_8y_9z_9y_{10}z_{10}$, which yields $D' = 11$.

We now repeat the experiment and select ten random numbers from the standard Normal distribution, that is, $N(0,1)$, and another ten from $N(1,1)$; the numbers are displayed to two places. The result is

$$N(0,1) \to \{x^{(1)}: -0.079,\ 1.3,\ -0.15,\ 1.2,\ 0.75,\ -1.2,\ -0.11,\ -0.84,\ 0.35,\ 0.55\}$$
$$N(1,1) \to \{x^{(2)}: 1.2,\ 0.056,\ 0.3,\ -0.77,\ 0.95,\ 1.1,\ 0.095,\ -0.43,\ 1.1,\ 1.3\}.$$

Here the ordered y sequence is associated with the $N(0,1)$ and the ordered z sequence is associated with $N(1,1)$. After combining the two sequences into a single sequence and ordering all the elements of this single sequence by size, we get the sequence

$$y_1y_2z_1z_2y_3y_4y_5z_3z_4z_5y_6y_7y_8z_6z_7z_8z_9y_9y_{10}z_{10},$$

which yields $D' = 8$ and has 27 percent fewer D's than in the $N(0,1)$. This example suggests that the RV D can be used as a statistic for testing the hypothesis that the populations are the same. If D is large enough, say $D > d_0$, we may conclude that the two samples come from the same population; else we reject that they come from the same population. The choice of d_0 is discussed below.

We test whether two samples come from the same population using the principles of hypothesis testing. We have two sets of samples: $\{X_i^{(1)}, i = 1, \ldots, n_1\}$ and $\{X_i^{(2)}, i = 1, \ldots, n_2\}$.

The null hypothesis, H_1, is that the two samples come from the same population, while the alternative, H_2, is that they do not come from the same population or, perhaps more accurately, that there is not enough evidence that they come from the same population. The test will be based on observing the test statistic D. If the event $\{D > d_0\}$ occurs, then the two samples interweave well and we may conclude that they come from the population. If the event $\{D \leq d_0\}$ occurs, we may conclude that H_1 is not supported by the data. If $\alpha \triangleq P[\text{rejecting } H_1|H_1 \text{ true}]$ denotes the level of significance, then $\alpha = P[D \leq d_0|H_1 \text{ true}] = \sum_{\text{all } d \leq d_0} P_D(d; n_1, n_2)$, where $P_D(d; n_1, n_2)$ is the probability of observing d runs in interwoven sequences of lengths n_1 and n_2.

Computing $P_D(d; n_1, n_2)$ requires some rather sophisticated counting procedures so we give only the final result here. Define

$$C_m^n \triangleq \binom{n}{m}.$$

Under the null hypothesis we find that

$$P_D(d; n_1, n_2) = \begin{cases} 2C_{(d/2)-1}^{n_1-1} C_{(d/2)-1}^{n_2-1} / C_{n_1}^{n_1+n_2}, d \text{ even} \\ (C_{(d-1)/2}^{n_1-1} C_{(d-3)/2}^{n_2-1} + C_{(d-3)/2}^{n_1-1} C_{(d-1)/2}^{n_2-1}) / C_{n_1}^{n_1+n_2}, d \text{ odd.} \end{cases}$$

These unwieldy formulas do not yield much for the purpose of analysis and require machine computation to evaluate α. However, it has been shown that for $n_1 \geq 10, n_2 \geq 10$, the distribution of D is well approximated by a Normal CDF with approximate mean and variance given by, respectively,

$$\mu_D \approx \frac{2n_1 n_2}{n_1 + n_2}, \; \sigma_D^2 \approx 4(n_1 + n_2)\left(\frac{n_1}{n_1 + n_2}\right)^2 \left(\frac{n_2}{n_1 + n_2}\right)^2.$$

Hence we approximate $\alpha = P[D \leq d_0|H_1 \text{ true}] = \sum_{\text{all } d \leq d_0} P(d; n_1, n_2)$ with

$$\alpha = \sum_{\text{all } d \leq d_0} P(d) \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_\alpha} \exp\left(-\frac{1}{2}x^2\right) dx, \; z_\alpha \triangleq \frac{d_0 - \mu_D}{\sigma_D}.$$

Example 7.5-8

(*run test on sameness of two populations*) We request two sets of ten random numbers from RANDOM.ORG from a population $N(1,1)$ and order these by size as

$$N(1,1) \to \{y^{(1)} : -1.4, -0.33, 0.40, 0.44, 0.70, 0.74, 1.3, 1.3, 1.7, 2.4\}$$
$$N(1,1) \to \{y^{(2)} : -0.67, -0.21, 0.38, 0.38, 0.51, 0.71, 1.4, 1.5, 2.0, 2.9\}.$$

For calibration we co-join these two sequences into a single sequence and order the elements of the sequence by size. We find that the realization $D'_{cal} = 12$. We then request a set of random numbers from an "unknown" Normal distribution and order these by size as

$$\{y^{(3)} : -3.8, -2.5, -0.13, 2.2, 2.8, 3.0, 3.8, 4.6, 5.5, 5.8\}.$$

After interleaving these by size with the $\{y^{(1)}\}$ sequence and counting the runs, we get $D' = 6$. We wish to test the hypothesis that the $\{y^{(1)}\}$ and $\{y^{(3)}\}$ sequences come from the same population at the 0.05 level of significance. We solve

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_{0.05}} \exp\left(-\frac{1}{2}x^2\right)dx = 0.05$$

and find that $z_{0.05} = -1.65$. For the given sample sizes we find that $\mu_D = 10, \sigma_D = \sqrt{5}$. Thus, $d_0 = \sigma_D z_{0.05} + \mu_D = 6.3$ and since $D' < d_0$ (barely), we reject the hypothesis that $\{y^{(3)}\}$ comes from a $N(1,1)$ population. Indeed, in this case, the $\{y^{(3)}\}$ sequence comes from a $N(1,3)$ population.

Ranking Test for Sameness of Two Populations

Another procedure for testing the sameness of two populations is the so-called *ranking test.* Assume that we have two continuous populations X and Y with respective distribution functions $F_X(x)$ and $F_Y(y)$. We wish to test the hypothesis $H_1 : F_X = F_Y$ versus the alternative $H_2 : F_X \neq F_Y$. We take n_1 samples from X and n_2 from Y, co-join them, and order them by size. Then we assign to each element of the sequence a number denoting its place in the ascending order; for example, the event $X_1 < Y_1 < X_2 < X_3 < Y_2 < Y_3 < Y_4$ would be designated as

$$\begin{matrix} X_1 & Y_1 & X_2 & X_3 & Y_2 & Y_3 & Y_4 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{matrix}.$$

The number associated with each element is its *rank*, and the Y sequence has ranks 2, 5, 6, and 7. Here $n_1 = 3$, $n_2 = 4$. The rank of the last element in the sequence is $n_1 + n_2$ and the rank of the first is 1. It is shown elsewhere that the RV

$$T \triangleq \sum_{Y\,sequence} ranks$$

is a suitable test statistic to test the hypothesis that $F_X(x) = F_Y(x)$, for all x. If T is too large or too small, the hypothesis is rejected. To test the hypothesis at a level of significance α, we need the distribution of T under the null hypothesis. It is shown elsewhere ([7-22] to [7-24]) that when $n_1 > 7, n_2 > 7$ (ideally we would want them larger), T is approximately distributed as $N(\mu_T, \sigma_T^2)$ with $\mu_T = n_2(n_1 + n_2 + 1)/2, \sigma_T^2 = n_1 n_2(n_1 + n_2 + 1)/12$. In the example above we find $\mu_T = 16, \sigma_T^2 = 8$.

Example 7.5-9

(*ranking test on sameness of two populations*) We use the $\{y^{(1)}\}$ and $\{y^{(3)}\}$ sequence of Example 7.5-8, co-join them, and assign ranks to the elements of the ascending sequence. For the elements of the $\{y^{(3)}\}$ sequence, the ranks are 1, 2, 5, 13, 15, 16, 17, 18, 19, and 20; their sum is 126, $\mu_T = 105$, and $\sigma_T = 13.23$. The hypothesis is that the two sequences come from the same population. At a level of significance $\alpha = 0.05$, we solve $F_T(x_{0.025}) = 0.025$

and get $x_{0.025} = -1.96$ so that the critical region is $\{T > 131\} \cup \{T < 79\}$. So we accept the hypothesis—*in error*—that the two sequences come from the same population. At a significance level $\alpha = 0.1$ the critical region is $\{T > 127\} \cup \{T < 87\}$. Marginally above $\alpha = 0.1$, the hypothesis is rejected.

SUMMARY

Hypothesis testing is a major branch of statistics that deals with decision making in a random (i.e., probabilistic) environment. In the beginning of this chapter we put ourselves in the mind of a surgeon who faced a difficult decision regarding whether to operate on one of his patients. By using all available prior information and seeking to minimize the average risk, we derived the Bayes decision rule, which—arguably—is the most rational approach to making decisions when available information is of the probabilistic kind rather than being categorical. The Bayes decision rule leads to a likelihood ration test (LRT).

The *prior probabilities* (sometimes called *a priori* probabilities) required in Bayes testing may not always be available in which case the threshold in the LRT for accepting/rejecting the hypothesis is determined not by minimizing the average risk but by the specified error probability α, which is the probability of rejecting the hypothesis based on observational data when in fact the hypothesis is true. In the case of testing a simple hypothesis versus a simple alternative, the Neyman–Pearson Theorem ensures that the LRT is optimum in that it is *the most powerful test.* By this is meant that the probability of rejecting the alternative hypothesis when it is true is driven to a minimum.

In a number of situations, testing a simple hypothesis versus a simple alternative won't do because the hypothesis or the alternative or both involve many outcomes in the underlying sample space. In that case the generalized likelihood ratio test (GLRT) is useful. We illustrated the GLRT with a number of examples and, in doing so, encountered such classic statistical tests as the F-test, the t-test, and the Pearson Chi-square test.

We then considered *ordering*, *percentiles*, and *rank* and illustrated how these tools can be made useful in *distribution-free* (sometimes called *robust*) statistics. We illustrated these with hypothesis testing examples using *run* tests and *ranking* tests.

PROBLEMS

7.1 Prove Equation 7.1-6.

7.2 Consider Example 7.1-1. Let the prior probabilities be $P_1 = 0.9, P_2 = 0.1$. How does this affect the Bayes decision rule?

7.3 Assume a Normal population $X{:}N(\mu, 1)$ and a sequence of i.i.d. observations on X, that is, $\{X_i : i = 1, \ldots, n\}$. Find the critical region for testing the hypothesis that $H_1 : \mu = \mu_1$ versus the alternative $H_1 : \mu > \mu_1$ at the 0.05 level.

7.4 Show that the power P of an LRT is given by $\text{P} = P[\text{reject } H_1 | H_2 \text{ is true}]$.

7.5 Why was it not necessary to invoke the Central Limit Theorem to argue that $\hat{\mu}_X(n)$ in Example 7.2-2 is Normally distributed?

7.6 We flip a coin 100 times and observe $50 + k$ heads and $50 - k$ tails. What is the largest value of k that will enable us to accept the hypothesis that the coin is fair at $\alpha = 0.05$ significance. Repeat for $\alpha = 0.01$.

7.7 A customer in a sub-freezing environment is considering buying an automobile battery at DBW ("Discount Battery Warehouse"). The particular battery model of interest is imported from one of two possible sources, say A and B, which do not share the same quality-control standards. The better import (A) will start the car 90 percent of the time in sub-freezing weather while the worse import (B) will start the car only 50 percent of the time in such weather. There are an equal numbers of batteries from each source. The imports cannot be differentiated by any external visible features. The battery salesman will allow the customer only one try at starting his car with a test battery, under sub-freezing conditions, before purchase.

We shall treat the customer's dilemma, such as it is, from a hypothesis testing point of view. Let the hypothesis be H_1: the battery start-probability $p_1 = 0.9$ versus the alternative H_2: the battery-start probability $p_2 = 0.5$. There are two actions: a_1 (buy the battery) and a_2 (reject the battery). The loss functions are in dollars: $l(a_1, p_1) = 0$; $l(a_1, p_2) = 40$ (money spent on a poor battery); $l(a_2, p_1) = 10$ (passing up a good deal that would cost at least \$10 elsewhere); $l(a_2, p_2) = 0$. Define the RV X as

$$X \triangleq \begin{cases} 1, & \text{if battery starts the car in test trial,} \\ 0, & \text{if battery fails to start car in test trial.} \end{cases}$$

(a) Define the four possible decision functions $(d_i, i = 1, \ldots, 4)$;
(b) Compute the risk for each decision function $(R(d_i; p_j), i = 1, \ldots, 4; j = 1, 2)$;
(c) Plot the risk function points in a Cartesian system where the abscissa is $R(d; p_1)$ and the ordinate is $R(d; p_2)$. From the graph, determine which decision function is *dominated* (is worse) by at least one other decision function and therefore is *inadmissible* (not worthy of consideration).
(d) Suppose it is known that there are twice as many batteries from import B as from A; how would this affect your decision?

7.8 Suppose a manufacturer of memory chips observes that the probability of chip failure is $p = 0.05$. A new procedure is introduced to improve the design of chips. To test this new procedure, 200 chips could be produced using this new procedure and tested. Let the random variable X denote the number of chips that fail out of these 200. We set the test rule that we would accept the new procedure if $X \leq 5$. Find the probability of a type I error.

7.9 Let $X{:}N(\mu, 1)$, where $\mu = \mu_1 = 1/2$ or $\mu = \mu_2 = -1/2$. Let $H_1 : \mu = -1/2$ and $H_2{:}\mu = 1/2$. Define the two actions a_1 : *accept* H_1(*reject* H_2) and a_2 : *accept* H_2 (*reject* H_1). The sample space for X is $\Omega = \{-\infty, \infty\}$. Let $S_1 = \{-\infty, 0\}$ and $S_2 = \{0, \infty\}$. Consider the two mutually exclusive events $E_1 = \{X \in S_1\}$ and $E_2 = \{X \in S_2\}$.

(a) Compute the four probabilities $P(E_i|\mu_j)i = 1, 2; j = 1, 2$;
(b) Define the four possible decision functions $d_i, i = 1, \ldots, 4$;
(c) Assuming the loss functions $l(a_1, \mu_1) = 0, l(a_1, \mu_2) = 2, l(a_2, \mu_1) = 5, l(a_2, \mu_2) = 0$, compute the risks associated with each of the decision functions in (b). Which decision function is *inadmissible*, that is, there is at least one other decision function that *dominates* (is better than) it?

7.10 We have two Normal populations $X_1 : N(\mu_1, \sigma^2)$ and $X_2 : N(\mu_2, \sigma^2)$. We test $H_1 : \mu_1 = \mu_2$ versus $H_2 : \mu_1 \neq \mu_2$ at a level of significance of 5 percent. Describe the test.

7.11 We have two Normal populations $X_1 : N(\mu_1, \sigma^2)$ and $X_2 : N(\mu_2, \sigma^2)$. We test $H_1 : \mu_1 = \mu_2$ versus $H_2 : \mu_1 > \mu_2$ at a level of significance of 5 percent. Describe the test.

7.12 Repeat Problem 7.11 with the change that $H_1 : \mu_1 = \mu_2$ versus $H_2 : \mu_1 < \mu_2$.

7.13 Suppose that we have n observations $X_i, i = 1, 2, \ldots, n$ of radar signals, and X_i are normal independently and identically distributed random variables. Under H_0, X_i have mean μ_0 and variance σ^2, while under H_1, X_i have mean μ_1 and variance σ^2, and $\mu_1 > \mu_0$. Determine the maximum likelihood test.

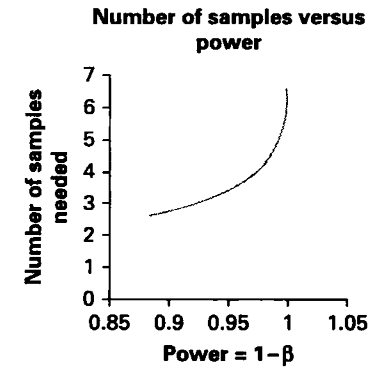
7.14 A manufacturer is interested in the output voltage of a power supply used in a personal computer. The output voltage is assumed to be normally distributed with a standard deviation of 0.25 volts, and the manufacturer wishes to test $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ using $n = 16$ units.

(a) The acceptance region is $4.85 \leq \bar{X} \leq 5.15$. Find the type I error.
(b) Find the power of the test for detecting a true mean output voltage of 5.1 volts.

7.15 Let $X : N(\mu, 1)$ represent a population whose mean is known to be $\mu = \mu_1 = 3$ or $\mu = \mu_2 = 1$. We make n i.i.d. observations on X and call these $\{X_i, i = 1, \ldots, n\}$. Let $H_1 : \mu = \mu_1 = 3$ and $H_2 : \mu = \mu_2 = 1$; show that the LRT is reduced to accept H_1 if $\hat{\mu} > (2n)^{-1} \ln(k) + 2 \triangleq c_n$, where, as usual, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$. The constant c_n is determined by the significance α. Find a general expression for c_n in terms of μ_1, n, and z_α, the latter being the α percentile of the $N(0, 1)$ distribution. Assuming n =10, what is the value of c_n for $\alpha = 0.01$?

7.16 (continuation of Problem 7.15) In Problem 7.15 treat n as an unknown and calculate the value of n needed to obtain $\alpha = 0.02$ and $\beta = 0.01$ simultaneously.

7.17 (continuation of Problem 7.16) Keeping α at $\alpha = 0.02$ show that the number of samples needed to achieve a given power follows the graph below. (Hint: Use NORMINV (probability, mean, standard deviation) in Excel ™.)



7.18 (F-test for comparing variances) The F-test is useful in testing whether the variances (or standard deviations) of two Normal populations are the same. Typically we test the hypothesis $H_1 : \sigma_1 = \sigma_2$ versus $H_2 : \sigma_1 \neq \sigma_2$. The F-test can be done online by entering the data from two Normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ and taking the ratio of the sample variances. Thus, assume we have m *samples* from population P1 $\{X_{1i}, i = 1, \ldots, m\}$ and n samples from population P2 $\{X_{2j}, j = 1, \ldots, n\}$. We do not mix the samples because it is important to keep the sample variances independent of each other. One of several programs will compute from the input realizations $\{x_{1i}, i = 1, \ldots, m\}$ and $\{x_{2j}, i = 1, \ldots, n\}$ the numerical sample variances, often denoted by the symbols s_1^2 and s_2^2, as $s_1^2 \triangleq (m-1)^{-1} \sum_{i=1}^{m} (x_{1i} - \bar{x}_1)^2$(degrees of freedom $DOF = m - 1$) and $s_2^2 \triangleq (n-1)^{-1} \sum_{i=1}^{n} (x_{2j} - \bar{x}_2)^2$(degrees of freedom $DOF = n - 1$). In these expressions $\bar{x}_1 = m^{-1} \sum_{i=1}^{m} x_{1i}$ and $\bar{x}_2 = n^{-1} \sum_{j=1}^{n} x_{2j}$ are the sample numerical means. We need to specify the significance level α. The algorithm then proceeds as follows: (1) compute $F' = s_1^2/s_2^2$; (2) compare F' with $F_{\alpha/2,\nu_1,\nu_2}$, where $F_{\alpha/2,\nu_1,\nu_2}$ is the critical value of the F-distribution with $m-1$ and $n-1$ degrees of freedom and significance α. When testing $H_1 : \sigma_1 = \sigma_2$ versus $H_2 : \sigma_1 > \sigma_2$ reject H_1 if $F' > F_{\alpha,\nu_1,\nu_2}$.

When testing $H_1 : \sigma_1 = \sigma_2$ versus $H_2 : \sigma_1 < \sigma_2$ reject H_1 if $F' < F_{1-\alpha,\nu_1,\nu_2}$.

When testing $H_1 : \sigma_1 = \sigma_2$ versus $H_2 : \sigma_1 \neq \sigma_2$ reject H_1 if $F < F_{1-\alpha/2,\nu_1,\nu_2}$ or $F > F_{\alpha/2,\nu_1,\nu_2}$.

As an exercise, generate two sets of Gaussian random numbers first with the same σ and then with different σ's and test the efficacy of the F-test using an online calculator, for example, the BioKin statistical calculator.

7.19 The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ boards has been collected and the following number of defects observed.

No.of defects	Observed frequencies
0	32
1	15
2	9
3	4

Test the goodness of fit.

7.20 A semi-conductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout of fraction defective at a critical manufacturing step does not exceed 0.05, and that the manufacturer demonstrate process capability at this level of quality using $\alpha = 0.05$. The semi-conductor manufacturer takes a random sample of 200 devices and finds that 4 of them are defective. Can the manufacturer demonstrate process capability for this customer?

7.21 (F-test) We are given the following factual data from [7-6] that tests the oxygen assimilating capability of various levels of smokers versus nonsmokers. There are five categories:

Category	Mean respiratory flow rate	Standard deviation of flow rate	Number of people in category
Nonsmokers in smoke-free environment (1)	3.17	0.74	200
Nonsmokers in smoky environment (2)	2.72	0.71	200
Light smokers (3)	2.63	0.73	200
Moderate smokers (4)	2.29	0.70	200
Heavy smokers (5)	2.19	0.72	200

The hypothesis H_1 is that there is no difference in air flow among the five categories; the alternative is that there is at least one category whose respiratory statistics are significantly different from the others.†
Compute whether to accept or reject the hypothesis at the 0.05 significance level.

7.22 (Chi-square test) Plant biologists attempt to test Mendel's law of hereditary by crossing two pea plants. According to Mendel's law three-fourth of the offspring should be green (dominant color) and one-fourth should be yellow (recessive). In 880 plants, the biologists observe 639 green seeds and 241 yellow seeds. Let H_1: green allele† is dominant and H_2: green allele is not dominant. Determine at the 0.05 level of significance whether to accept or reject the hypothesis.

7.23 Let $(X_1, X_2, \ldots, X_n)$ be a random sample of a normal random variable with mean μ and variance 100. Let
$H_0 : \mu = 50$
$H_1 : \mu = \mu_1 (> 50)$
and sample size n=25. As a decision procedure, we use the rule to reject H_0 if $\bar{x} \geq 52$, where $\bar{x}$ is the value of the sample mean $\bar{X}$.

(a) Find the probability of rejecting H_0: $\mu = 50$ as a function of $\mu(> 50)$.
(b) Find the probability α of a type I error.
(c) Find the probability β of a type II error when (i) $\mu_1 = 53$ and (ii) $\mu_1 = 55$.

†Note that if we reject the hypothesis, we still won't know which category (or categories) was responsible for the rejection.

†A gene transferring inherited characteristics.

7.24 Show that the statistic V in the Pearson goodness-of-fit test has expectation $E[V|H_1] = l-1$ under hypothesis H_1.

7.25 Show that the statistic V in the Pearson goodness-of-fit test has expectation $E[V|H_2] > l-1$ under the alternative H_2.

7.26 Consider the F-test in testing for the equality of two variances. Plot the test statistic versus the variance ratio for $m = 8, n = 5$. Find the critical region for significance of 0.05.

7.27 In testing the equality of two variances of two Normal populations with m samples from population P1 and n samples from population P2, show that when H_1 is true Λ can be written as

$$\Lambda = A(m,n)\frac{\left(\frac{(m-1)}{(n-1)}F_{m-1,n-1}\right)^{m/2}}{\left(1+\frac{(m-1)}{(n-1)}F_{m-1,n-1}\right)^{(m+n)/2}},$$

where $A(m,n) \triangleq (m+n)^{(m+n)/2}m^{-m/2}n^{-n/2}$.

7.28 Aircrew escape systems are powered by a solid propellant. Specifications require that the mean burning rate must be 50 cm per second. The standard deviation of burning rate is $\sigma = 2$ cm per second. A random sample of n=25 is obtained and the sample burning rate $x'' = 51.3$ cm per second is calculated. What conclusion should be drawn at a significance level of $\alpha = 0.05$?

7.29 A melting point test of $n = 10$ samples of a binder used in manufacturing a rocket propellant resulted in $\bar{x} = 154.2\,^\circ F$. Assume that the melting point is normally distributed with $\sigma = 1.5\,^\circ F$.

(a) Test $H_0\colon \mu = 155$ versus $H_0\colon \mu \neq 155$ using $\alpha = 0.01$.
(b) Calculate the power of the test if true mean is $\mu = 150$.

7.31 Twenty-four observations are made on a random variable X and are ordered by size as $Y_1 < Y_2 < \cdots < Y_{24}$. Estimate the 30th percentile.

7.32 Find a 98 percent confidence interval for the median from 25 samples.

7.33 The mean lifetime of a sample of 100 lightbulbs produced by Lighting Corporation is computed to be 1570 hours with a standard deviation of 120 hours. If the president of the company claims that the mean lifetime $E[X]$ of all the lightbulbs produced by the company is 1600 hours, test the hypothesis that $E[X]$ is not equal to 1600 hours using a level of significance of (a) 0.05 and (b) 0.01.

7.34 A manufacturer of a migraine headache drug claimed that the drug is 90% effective in relieving migraines for a period of 24 hours. In a sample of 200 people who have migraine headaches, the drug provided relief for 160 people for a period of 24 hours. Determine whether the manufacturer's claim is legitimate at a level of significance of 0.05.

REFERENCES

General

7-1. A. M. Mood and F. A. Graybill, *Introduction to the Theory of Statistics*, 2nd edition. New York: McGraw-Hill, 1963.
7-2. A. Papoulis, *Probability & Statistics.* Englewood Cliffs, NJ: Prentice Hall, 1990.
7-3. R. Walpole, R. Myers , and K. Ye, *Probability and Statistics for Engineers and Scientists*, 7th edition. Delhi: Pearson Education, 2002.
7-4. A. L. Garcia, *Probability Statistics and Random Processes for Electrical Engineering,* 3rd edition. Upper Saddle River, NJ: Prentice Hall, 2008

F-test

7-5. (Online) BioKin at: http://www.biokin.com/tools/fcrit.html
7-6. S. A. Glanz, *A Primer on Biostatistics*, 3rd edition. McGraw-Hill, New York, 1992.
7-7. N. J. Solkind, *Statistics for People Who (Think They) Hate Statistics,* 2nd edition. Sage Publications, 2010.
7-8. G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 8th edition. Ames, IA: Iowa State Press, 1989.
7-9. (Online) NISTI/SEMATECH *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, date

T-test

7-10. (Online) "The T-Test," available at http:// www.socialresearchmethods.net/kb/stat_t.php
7-11. D. W. Zimmerman, "A Note on Interpretation of the Paired-Sample t-Test," *Journal of Educational and Behavioral Statistics*, Vol. 22, No. 3, pp. 349–360, 1997.
7-12. Student (W. S. Gossett), "The Probable Error of the Mean," *Biometrica,* Vol. 6, No.1, pp. 1–25, 1908.
7-13. S. J. Coakes and L. G. Steed, *SPSS: Analysis for Windows: Version 7.0, 7.5, 8.0 for Windows*, Wiley, Brisbane, Australia, 1999.
7-14. H. B. Mann and D. R. Whitney, "On a Test Whether One of Two Random Variables is Stochastically Larger Than the Other," *Annals of Mathematical Statistics*, Vol. 18, pp. 50–60, 1947.
7-15. (Online) Student's t-distribution, available at http://en.wikipedia.org/wiki/student's_t-distribution

Chi-square test

7-16. Chernoff and E. L. Lehmann, "The Use of maximum Likelihood Estimates in χ^2 tests for Goodness-of-fit," *The Annals of Mathematical Statistics*, Vol. 25, pp. 579–586, 1954.
7-17. R. L. Plackett, "Karl Pearson and the Chi-Square Test," *International Statistical Institute (ISI)*, Vol. 51, No. 1, pp. 59–72, 1983.

Pearson test

7-18. K. Pearson, "On the Criterion that a Given System of Deviations from the Probable... Have Arisen from Random Sampling," *Philosophical magazine*, Series 550, Vol. 302, pp. 157–175, 1900.

7-19. J. Neyman and E. S. Pearson, "On the Use and Interpolation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrica*, Vol. 20, pp. 175–240, 1928.

7-20. J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, Series A, Vol 231, pp. 289–337, 1933.

7-21. O. Zeitouni, J. Ziv, and N. Mershav, "When is the GLRT Optimal?" IEEE Transactions on Information Theory, Vol.38, pp.1597-1601,1991.

Order statistics and ranking

7-22. F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics,* Vol 1, pp. 80–83, 1945.

7-23. S. S. Wilks, "Order Statistics," *Bulletin of the American Mathematical Society*, Vol 54, pp. 6–50, 1948.

7-24. S Siegel, *Non-parametric Statistics for Non-Statisticians: A Step-by-Step Approach*, Upper Saddle River, NJ: Wiley, 2009.

8 Random Sequences

Random sequences are used as models of sampled data arising in signal and image processing, digital control, and communications. They also arise as inherently discrete data such as economic variables, the content of a register in a digital computer, something as simple as coin flipping (Bernoulli trials), or the number of packets on a link in a computer network. In each case, the random sequence models the unpredictable behavior of these sources from the user's perspective. In this chapter we will study the random sequence and some of its important properties. As we will see, a random (stochastic) sequence can be thought of as an infinite dimensional vector of random variables.† As such it stands between finite dimensional random vectors (cf. Chapter 5) and continuous-time random functions, called random processes, to be studied in the next chapter.

Another way to generalize the random vector is by doubling the number of index parameters to two, thereby creating random matrices, which have been found useful as mathematical models in image processing. When these random matrices grow in size, in the infinite limit we have a two-dimensional random sequence, used in many theoretical studies in image and geophysical signal processing. While we will not study image processing here, many of the basic concepts of random sequences carry over to the two-dimensional case. Three- and four-dimensional random sequences have been found useful models of unpredictable aspects in video and other spatiotemporal signals.

†In the real world all sequences are finite. However, as long as the real-world sequences are long compared to internal correlations, the infinite length model does not significantly detract from accuracy except when we are at the very beginning or end of the real-world sequence.

8.1 BASIC CONCEPTS

In the course of developing this material we will have need to review and extend some of the basic material presented in Chapter 1 on the axioms of probability. This is because we must now routinely deal with an infinite number of random variables at one time, that is, a random sequence. We start out this study by offering a definition of the random sequence followed by a few simple examples.

Definition 8.1-1 Let $(\Omega, \mathscr{F}, P)$ be a probability space. Let $\zeta \in \Omega$. Let $X[n, \zeta]$ be a mapping of the sample space Ω into a space of complex-valued sequences on some index set Z. If, for each fixed integer $n \in Z$, $X[n, \zeta]$ is a random variable, then $X[n, \zeta]$ is a *random (stochastic) sequence*. The index set Z is *all* the integers, $-\infty < n < +\infty$, padded with zeros if necessary. ■

See Figure 8.1-1 for an illustration for sample space $\Omega = \{1, \ldots, 10\}$. We see that $X[n, \zeta]$ for a *fixed* outcome ζ is an ordinary sequence of numbers, that is, a deterministic (nonrandom) function of the discrete parameter n. We often refer to these ordinary sequences as *realizations* of the random sequence, or as *sample sequences* and denote them by $X_\zeta[n]$ or merely by $x[n]$ when there is no confusion. Thus, ten sample sequences are plotted in Figure 8.1-1, one for each outcome $\zeta \in \Omega$. On the other hand, for n *fixed* and ζ variable, $X[n, \zeta]$ is a random variable.† Thus the *collection* of all these realizations, $-\infty < n < +\infty$, along with the probability space, *is* the random sequence. We shall often, but not always,

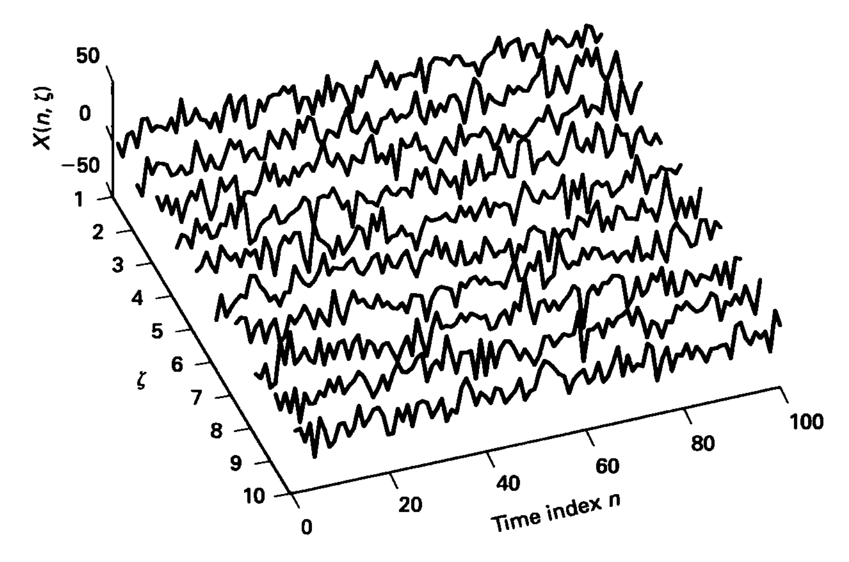


Figure 8.1-1 Illustration of the concept of *random sequence* $X(n, \zeta)$, where the ζ domain (i.e., the sample space Ω) consists of just ten values. (Samples connected only for plot.)

†Elementary probability texts talk about an i.i.d. sequence of RVs denoted by $X_n(\zeta)$. Our random sequence however, allows the added complication of dependence among these RVs.

denote the random sequence by just $X[n]$. We retain the notation $X[n,\zeta]$ when its use helps to clarify a point on the outcomes ζ of the underlying sample space Ω. Note that we use square brackets around the time argument n here, as is the convention in discrete-time signal processing.

We give the following simple examples of random sequences:

Example 8.1-1

(*separable random sequence*) Let $X[n,\zeta] \triangleq X(\zeta)f[n]$, where $X(\zeta)$ is a random variable and $f[n]$ is a given deterministic (ordinary) sequence. Such a random sequence is the separable product of a random variable (function) and an ordinary sequence. We will also write $X[n] = Xf[n]$, suppressing the outcome ζ variable, as is the custom for random variables. We see that all the sample sequences are just scaled versions of one another, with the scalar being the random variable X.

Example 8.1-2

(*sinusoid with random amplitude and phase*) Let $X[n,\zeta] \triangleq A(\zeta)\sin(\pi n/10 + \Theta(\zeta))$, where A and Θ are random variables defined on a common probability space $(\Omega, \mathscr{F}, P)$, alternately written $X[n] = A\sin(\pi n/10 + \Theta)$.

These two simple random sequences are made from deterministic components, but they are also "deterministic" in another way. They have the unusual property, from a probabilistic standpoint, that their future values are exactly determined from their present and past values. In Example 8.1-1, once we observe $X[n]$ at any fixed value of n, say $n = 0$, then, since the ordinary sequence $f[n]$ is assumed to be known and nonrandom, all of the random sequence $X[n]$ becomes known. We see that the random sequence $X[n]$ is *conditionally known* given its value at $n = 0$. The situation in Example 8.1-2 is just slightly more complicated but the same approach suffices to show that given two (nondegenerate) observations, say at $n = 0$ and $n = 5$, one can determine the values taken on by the random variables A and Θ; then the sequence $X[n]$ becomes conditionally known or perfectly predictable given these observations at $n = 0$ and $n = 5$. These deterministic random sequences would not be good models for noise on a communications channel because real noise is not so easily foiled.

In the next example we see how a more general but still "deterministic" random sequence can be made out of a random vector.

Example 8.1-3

(*random sequence with finite support*) Let $X[n,\zeta]$ be given by

$$X[n,\zeta] \triangleq \begin{cases} X_n(\zeta), & 1 \le n \le N, \\ 0, & \text{else}. \end{cases}$$

Since $X[n] = 0$ except for $n \in [1, N]$, we say $X[n]$ has *finite support*. Because of this finite support property, we can model this random sequence by a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_N)^T$ and then use the rich calculus of matrix algebra, for example, covariance matrices and linear transformations, as presented in Chapter 5. Many random sequences can be approximated this way, although note that we would have to consider the limiting behavior of such $\mathbf{X}$, as $N \to \infty$, to model a general random sequence.

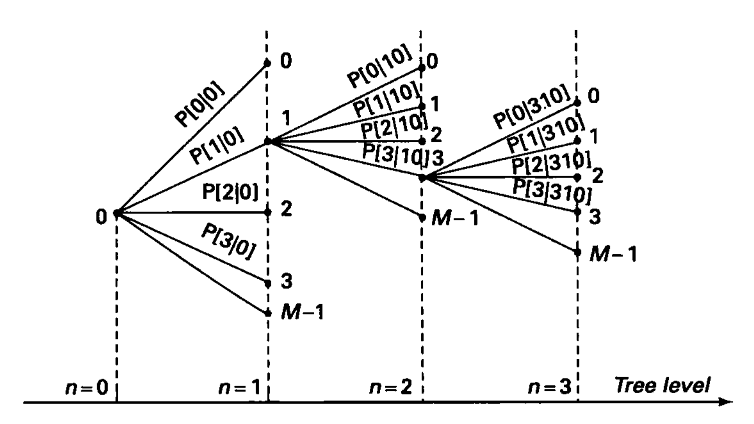


Figure 8.1-2 Tree diagram for discrete amplitude random sequence.

Example 8.1-4

(*tree diagram for random sequence*) Let the random sequence $X[n]$ be defined over $n \geq 0$, and take on only M discrete values, $0, 1, 2, \ldots, M-1$. Further assume the starting value is pinned at $X[0] = 0$. Then we can illustrate the evolution of the sample sequences of this random sequence with a tree diagram, with branching factor M at each node $n = 0, 1, 2, \ldots$ as illustrated in Figure 8.1-2.

At each level n, of the tree, the node values give possible sample sequence values $x[n]$, with branch index $i = 0, \ldots, M-1$. The sample sequences are identified by the sequence of node values of a path through the tree starting from the root node $n = 0$. If we identify the path string $i_1 i_2 i_3 \ldots$ with the base-M number $0.i_1 i_2 i_3 \ldots$, we can call this point the outcome $\zeta \in [0, 1] = \Omega$, the sample space.† Finally we can label the branches with the conditional probability $P[X[n] = m_i | \{X[k] \text{ on same path for } k \leq n-1\}]$, which in Figure 8.1-2 is denoted as $P[i_n | i_{n-1} i_{n-2} \ldots i_1 0]$. Then the probability of any node value at tree level n is just given by the product of all the probability branch labels back to the root node along this path. Note that all sample sequences that agree up to time n will correspond to a neighborhood in the sample space $\Omega = [0, 1]$ of radius $\frac{1}{2} M^{-n}$.

This example also has shown how to construct a consistent underlying sample space in the common case where we are given just the probability distribution information about the set of random variables that make up the random sequence. Note that when the random variables are all independent of one another, that is, jointly independent, and this probability distribution doesn't change with time, the branch labels in the tree are all the same, and in effect, the tree collapses to one stage. This is the situation called a sequence of i.i.d. random variables in probability theory. Generalizing this slightly we have the following definition.

Definition 8.1-2 An *independent random sequence* is one whose random variables at any time $n_1, n_2, \ldots, n_N$ are jointly independent for all positive integers N. ■

†For example let $M = 8$ and consider the base 8 number $0.1200\ldots0\ldots$. This implies that $X[1] = 1, X[2] = 2$, and all subsequent values are 0.

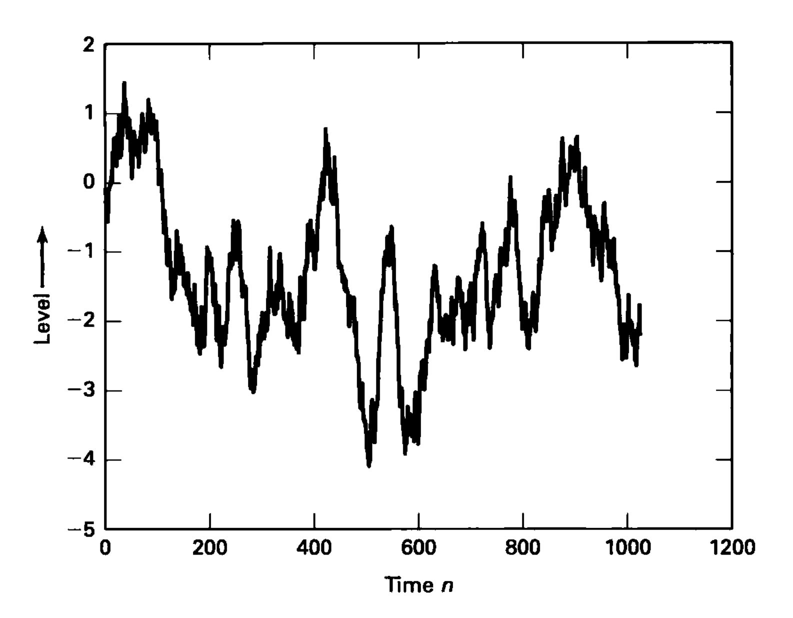


Figure 8.1-3 Example of a sample sequence of a random sequence. (Samples connected only for plot.)

Independent random sequences play a key role in our theory because they are relatively easy to analyze, they form the basis of more complicated and accurate models, and it is easy to get approximate sample sequences using random number generators on computers. Also when the discrete data arises by sampling continuous-time data, statistical independence often is a good approximation if the samples are far apart.

Figure 8.1-3 shows a segment from a real noise sequence, and Figure 8.1-4 shows a close-up portion revealing its discrete-time nature and detailed "randomness." This segment could have been taken from anywhere in the noise sequence and the statistical properties would have been the same. This remarkable property hints at some form of "stationarity" which will shortly be defined (Definition 8.1-5). Note that successive random variables, making up this segment, do not appear to be independent. Rather they are evidently correlated, necessitating in general an Nth-order probability distribution to statistically describe just this segment of this noise sequence. Continuing in this way, we would need an infinite-order CDF to characterize the whole random sequence!

In order to deal with infinite length random sequences, we may have to be able to compute the probabilities of *infinite intersections*† of events, for example, the event $\{X[n] < 5 \text{ for all positive } n\}$, which can be written as either $\bigcap_{n=1}^{+\infty}\{X[n] < 5\}$ or, by De Morgan's laws, in terms of the infinite union $\left(\bigcup_{n=1}^{\infty}\{X[n] \geq 5\}\right)^c$. This requires that we can *define* and work with the probabilities of infinite collections of events, which presents a problem with Axiom 3 of probability measure: That is, for $AB = \phi$ the null set,

$$P[A \cup B] = P[A] + P[B] \quad \text{(Axiom 3)}. \tag{8.1-1}$$

†Please review Section 1.4 on the definition of infinite intersections and unions. The concept is simple but often misunderstood.

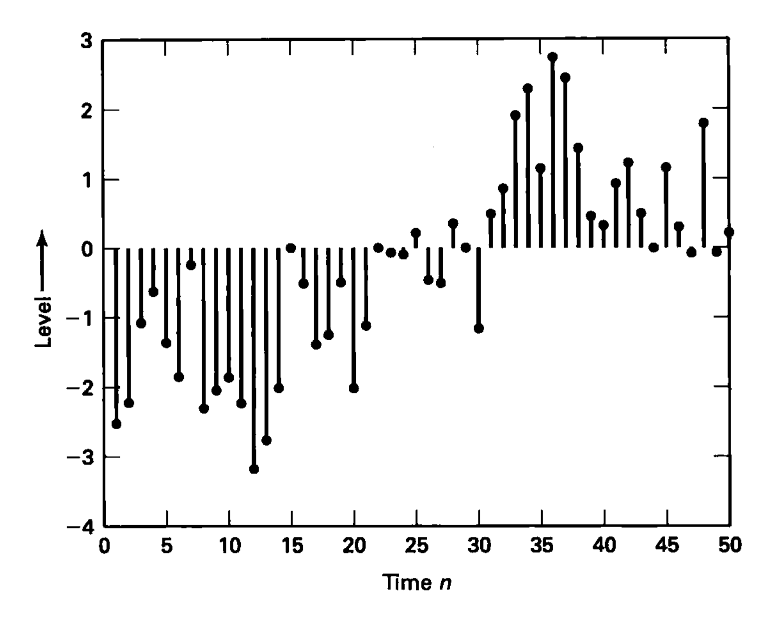


Figure 8.1-4 Close-up view of portion of sample sequence.

By iteration we could build this result up to the result

$$P\left[\bigcup_{n=1}^{N} A_n\right] = \sum_{n=1}^{N} P[A_n],$$

for any finite positive N, assuming $A_iA_j = \phi$ for all $i \neq j$. This is called *finite additivity.* It will permit us to evaluate $\lim_{N\to\infty} P[\bigcup_{n=1}^{N} A_n]$, but what we need above is $P[\bigcup_{n=1}^{\infty} A_n]$, where $A_n \triangleq \{X[n] \geq 5\}$. For general functions these two quantities might not be the same, that is, $\lim_{N\to\infty} f(x_N) \neq f(\lim_{N\to\infty} x_N)$. For this interchange of limiting operations to be valid, we need some kind of continuity built into probability measure P. This can be achieved by augmenting or replacing Axiom 3 by the stronger infinitely (countably) additive Axiom 4 given as

Axiom 4 (Countable Additivity)

$$P\left[\bigcup_{n=1}^{\infty} A_n\right] = \sum_{n=1}^{+\infty} P[A_n], \tag{8.1-2}$$

for an infinite collection of events satisfying $A_iA_j = \phi$ for $i \neq j$. ■

Fortunately, in the branch of mathematics called *measure theory* [8-1] (see also Appendix D), it is shown that it is always possible to construct probability measures satisfying the stronger Axiom 4. Moreover, if one has defined a probability measure P satisfying Axiom 3, that is, it is finitely additive, then the Russian mathematician Kolmogorov [8-2],

often referred to as the father of modern probability, has shown that it is always possible to extend the measure P to satisfy the countable additivity Axiom 4. We pause now for an example, after which we will show that Axiom 4 is equivalent to the desired continuity of the probability measure P. Henceforth, we will assume that our probability measures satisfy Axiom 4, and say they are *countably additive*.

Infinite-length Bernoulli Trials

Let $\Omega = \{\text{H,T}\}$, i.e. two outcomes $\zeta = \text{H}$ and T, with $P[\text{H}] = p$, with $0 < p < 1$, and $P[\text{T}] = q \triangleq 1 - p$. Define the random variable W by $W(\text{H}) \triangleq 1$ and $W(\text{T}) \triangleq 0$, indicative of successes and failures in coin flipping.

Let Ω_n be the sample space on the nth flip (the nth copy of Ω) and define a new event space as the infinite cross product† $\mathbf{\Omega}_\infty \triangleq \times_{n=1}^{\infty}\Omega_n$. This would be the sample space associated with an infinite sequence of flips, each with sample space Ω_n. We then define the random sequence $W[n, \boldsymbol{\zeta}] \triangleq W(\zeta_n)$, thus generating the *Bernoulli random sequence* $W[n]$, $n \geq 1$. Here the outcome $\boldsymbol{\zeta}$ is given as the outcomes at the individual trials as $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, ..., \zeta_n, ...)$.

Consider the probability measure for the infinite dimensional sample space $\mathbf{\Omega}_\infty$. Letting A_n denote an event‡ at trial n, that is, $A_n \in \mathscr{F}_n$, where $\mathscr{F}_n$ is the field of events in the probability space $(\Omega_n, \mathscr{F}_n, P)$ of trial n, we need to have $\bigcap_{n=1}^{\infty} A_n$ as an event in $\mathscr{F}_\infty$, the σ-field of events in $\mathbf{\Omega}_\infty$. To complete this field of events, we will have to augment it with all the countable intersections and unions of such events. For example, we may want to calculate the probability of the event

$$\{W[1] = 1, W[2] = 0\} \cup \{W[1] = 0, W[2] = 1\},$$

which can be interpreted as the union of two events of the form $\bigcap_{n=1}^{\infty} A_n$; that is, $\{W[1] = 1, W[2] = 0\} = \bigcap_{n=1}^{\infty} A_n$ with $A_1 = \{W[1] = 1\}$, $A_2 = \{W[2] = 0\}$, and $A_n = \Omega_n$ for $n \geq 3$. Hence $\mathscr{F}_\infty$ must include all such events for completeness. To construct a probability measure on $\mathbf{\Omega}_\infty$, we start with sets of the form $\boldsymbol{A}_\infty = \bigcap_{n=1}^{\infty} A_n$ and define *in the case of independent trials*,

$$\mathbf{P}_\infty[\mathbf{A}_\infty] \triangleq \prod_{n=1}^{\infty} P[A_n].$$

We then extend this probability measure to all of $\mathscr{F}_\infty$ by using Axiom 4 and the fact that every member of $\mathscr{F}_\infty$ is expressible as the countable union and intersection of events of the form $\bigcap_{n=1}^{\infty} A_n$.We have in principle thus constructed the probability space $(\mathbf{\Omega}_\infty, \mathscr{F}_\infty, \boldsymbol{P}_\infty)$ corresponding to the infinite-length Bernoulli trials, with associated Bernoulli random sequence

$$W[n, \boldsymbol{\zeta}] = W(\zeta_n), \qquad n \geq 1.$$

†Here the infinite cross product $\times_{n=1}^{\infty}\Omega_n$ simply means that the points in $\mathbf{\Omega}_\infty$ consist of all the infinite-length sequences of events, each one in Ω_n for some n. Thus if outcome $\boldsymbol{\zeta} \in \mathbf{\Omega}_\infty$, then $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3, \ldots)$, where outcome ζ_n is in Ω_n for each $n \geq 1$. (The finite-length case of Bernoulli trials was treated in Section 1.9.)

‡Most likely just a singleton event, that is, just one outcome, in this binary case.

We have just seen how to construct the sample space $\mathbf{\Omega}_\infty$ for the (infinite-length) Bernoulli random sequence, where the outcomes $\boldsymbol{\zeta}$ are just infinite-length sequences of "H" and "T." This $W[n]$ is thus our first nontrivial example of a random sequence. However, it may seem a bit artificial to regard each random variable $W[n, \boldsymbol{\zeta}]$ as a function of the infinite dimensional outcome vectors that make up the elements in the sample space $\mathbf{\Omega}_\infty$. It seems as though we have unnecessarily complicated the situation, after all $W[n, \boldsymbol{\zeta}]$ is just $W(\zeta_n)$. To see that this notational complication is unavoidable, let us turn to the commonly occurring model for correlated noise,

$$X[n] = \sum_{m=1}^{n} \alpha^{n-m} W[m], \text{ for } n \geq 1, \tag{8.1-3}$$

where $W[n]$ is the Bernoulli random sequence just created. Writing the filtered output $X[n]$ for each outcome $\boldsymbol{\zeta}$,

$$X[n, \boldsymbol{\zeta}] = \sum_{m=1}^{n} \alpha^{n-m} W(\zeta_m),$$

we see that each $X[n, \boldsymbol{\zeta}]$ is a function of an ever-increasing (with n) number of components of $\boldsymbol{\zeta}$, that is, the value of $X[n]$ depends on outcomes $\zeta_1, \zeta_2, \ldots, \zeta_n$. If we just dealt with each fixed value of n as a separate problem, that is, a separate sample space and probability measure, there would be the unanswered question of consistency. This is where, in practice, we would call on Kolmogorov's consistency theorem to show that our results are consistent with *one* sample space $\mathbf{\Omega}_\infty$ which has (infinite-length) outcomes $\boldsymbol{\zeta}$.†

Example 8.1-5

(*correlated noise*) Consider the random sequence in Equation 8.1-3, with $|\alpha| < 1$. We take the Bernoulli random sequence $W[n]$ as input, that is, $W[n] = 1$ with probability p, and $W[n] = 0$ with probability $q \triangleq 1 - p$. We want to find the mean of $X[n]$ at each positive n. Since the expectation operator is linear, we can write

$$\begin{aligned} E\{X[n]\} &= E\left\{ \sum_{m=1}^{n} \alpha^{n-m} W[m] \right\} \\ &= \sum_{m=1}^{n} \alpha^{n-m} E\{W[m]\} \\ &= \sum_{m=1}^{n} \alpha^{n-m} p = p \sum_{m=1}^{n} \alpha^{n-m} \\ &= p \sum_{m'=0}^{n-1} \alpha^{m'} = p \frac{(1-\alpha^n)}{(1-\alpha)}. \end{aligned}$$

†The use of bold notation for $\mathbf{\Omega}_\infty, \boldsymbol{\zeta}, \mathbf{P}_\infty$ is rather extravagant but was introduced to avoid confusion. Clearly, $\mathbf{\Omega}_\infty$ is not the same as $\lim_{n\to\infty} \Omega_n$. Each outcome in Ω_n is either a {H} or a {T} no matter how large n gets. On the other hand, the outcomes in $\boldsymbol{\zeta} \in \mathbf{\Omega}_\infty$ are infinitely long strings of H's and T's. In the future we shall dispense with the bold notation even if Ω is generated by an infinite cross product and its elements (outcomes) are infinitely long strings.

The random sequence $X[n]$ thus created is not a sequence of independent random variables, as we can see by calculating the correlation

$$\begin{aligned} E\{X[2]X[1]\} &= E\{(\alpha W[1] + W[2])\, W[1]\} \\ &= \alpha E\{W^2[1]\} + E\{W[2]\}E\{W[1]\} \\ &= \alpha p + p^2 \\ &\neq (\alpha + 1)p^2 = E\{X[2]\}E\{X[1]\}. \end{aligned}$$

The random variables $X[2]$ and $X[1]$ must be dependent, since they are not even uncorrelated.

However, since the $W[n]$ are uncorrelated we can easily calculate the variance $\text{Var}\{X[n]\}$ as

$$\begin{aligned} \text{Var}\{X[n]\} &= \sum_{m=1}^{n} \text{Var}\{\alpha^{n-m} W[m]\} \\ &= \sum_{m=1}^{n} \alpha^{2(n-m)} \text{Var}\{W[m]\} \\ &= \frac{(1-\alpha^{2n})}{(1-\alpha^2)} pq. \end{aligned}$$

The dynamics of this random sequence can be modeled using a difference equation. Since $X[n-1] = \sum_{m=1}^{n-1} \alpha^{n-1-m} W[m]$, it follows that $X[n] = \alpha X[n-1] + W[n]$, a result that clearly exhibits the dependence of $X[n]$ on its immediate neighbor $X[n-1]$.† Thus, correlated noise $X[n]$ can be generated from the independent sequence $W[n]$ by *filtering* with the configuration shown in Figure 8.1-5. From Equation 8.1-3 we see that for large n, $X[n]$ is the sum of a large number of independent random variables. Hence by the Central Limit Theorem it will tend to a Gaussian distribution, $n \to \infty$, with mean $p\frac{1-\alpha^n}{1-\alpha}$ and variance $pq\frac{1-\alpha^{2n}}{1-\alpha^2}$.

Zero-mean, correlated, Gaussian noise can be generated using the same model. Thus, with $W[1], W[2], \ldots, W[n], \ldots$ denoting zero-mean, independent, identically distributed,

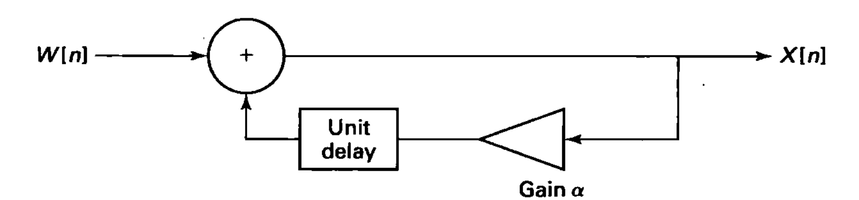


Figure 8.1-5 A feedback filter that generates correlated noise $X[n]$ from an uncorrelated sequence $W[n]$.

---

†Such explicit dependence in the equation like this is sometimes called *direct dependence.*

Gaussian random variables with $N(0, \sigma_W^2)$, the random sequence $X[n] = \sum_{m=1}^{n} \alpha^{n-m} W[m]$ will be zero-mean, Gaussian with variance

$$\mathrm{Var}\{X[n]\} = \frac{1-\alpha^{2n}}{1-\alpha^2}\sigma_W^2,$$

where $\sigma_W^2 = \mathrm{Var}\{W[n]\}$. Here too, the sequence produced by the filter is correlated since $E\{X[2]X[1]\} = \alpha E\{W^2[1]\} = \alpha\sigma_W^2 \neq E\{X[2]\}E\{X[1]\} = 0$.

The next example gives a MATLAB method to construct realizations of the Bernoulli random sequence and then passes the resulting sample sequences through a first-order filter to generate sample sequences of a (more realistic) correlated random sequence.

Example 8.1-6

(*sample sequence construction*) We use MATLAB to construct a sample sequence of $W[n]$. The MATLAB program

```
u = rand(40,1);
w = 0.5 >= u;
stem (w),
```

uses the built-in function "rand" to generate a 40-element vector of uniform random variables. The second line sets the vector elements $w[n]$ to 1 if $u[n] \geq 0.5$, and to 0 if $u[n] < 0.5$. So $w[n]$ is a sample sequence of the Bernoulli random sequence with $p = 0.5$. The corresponding MATLAB plot is shown in Figure 8.1-6.

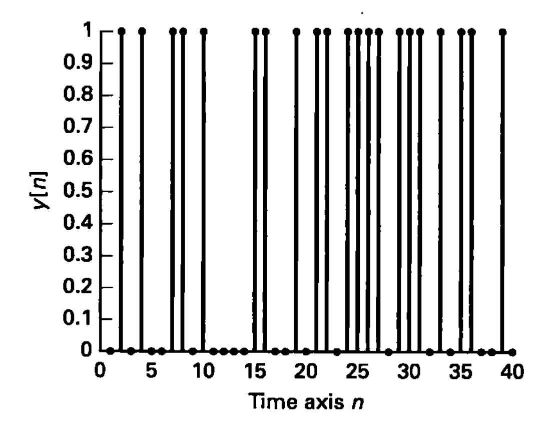


Figure 8.1-6 A sample sequence $w[n]$ for the Bernoulli random sequence $W[n]$.

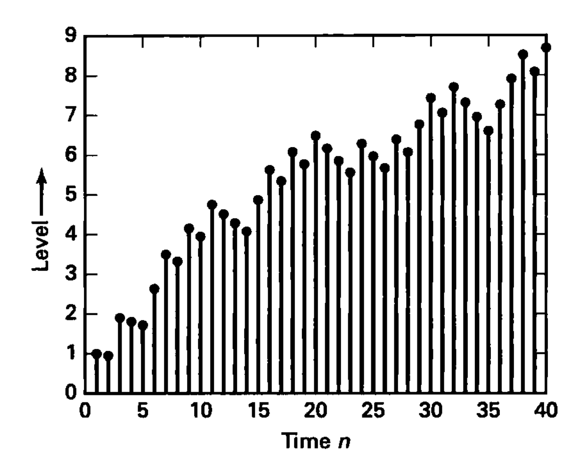


Figure 8.1-7 First 40 points illustrating startup transient.

To model the sample sequences of $X[n]$, which we denote $x[n]$, we can filter the sequence $w[n]$ with the filter,

$$x[n] = \alpha x[n-1] + w[n],$$

which has impulse response $h[n] = \alpha^n u[n]$ to realize the linear operation of Equation 8.1-3. The corresponding MATLAB m-file fragment is

```
b = 1.0;
a = [1.0 -alpha];
x = filter(b,a,w);
stem (x)
```

The result for $\alpha = 0.95$ and a 400-element vector was computed. Figure 8.1-7 shows the startup transient for the first 40 values. Figure 8.1-8 shows a sample of the approximate steady-state behavior starting at $n = 350$ and plotted for 50 points. Note the sample average value that has built up in $x[n]$ over time.

Note that the random sequence $X[n]$ has typical noise-like characteristics. The filter has correlated the random variables making up $X[n]$ so that sample sequences $x[n]$ look more "continuous." This simple example is called an *autoregressive* (AR) model and is widely used in signal processing to model both noises and signals. Note that the deterministic defect of the initial examples has now been removed. The reason is that the Bernoulli input sequence provides a new independent value for every sample, ensuring that the next sample cannot be perfectly predicted from the past.

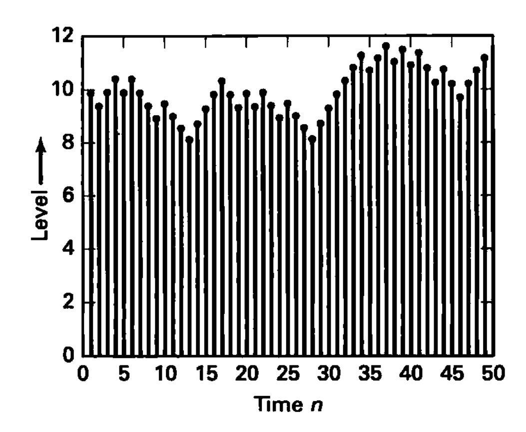


Figure 8.1-8 A segment of 50 points starting at $n = 350$.

Continuity of Probability Measure

When dealing with an infinite number of events, we have seen that continuity of the probability measure can be quite useful. Fortunately, the desired continuity is a direct consequence of the extended Axiom 4 on countable additivity (cf. Equation 8.1-2).

Theorem 8.1-1 Consider an *increasing sequence* of events B_n, that is, $B_n \subset B_{n+1}$ for all $n \geq 1$ as shown in Figure 8.1-9. Define $B_\infty \triangleq \bigcup_{n=1}^{\infty} B_n$; then $\lim_{n\to\infty} P[B_n] = P[B_\infty]$.

Proof Define the sequence of events A_n as follows:

$$A_1 \triangleq B_1$$

$$A_n \triangleq B_n B_{n-1}^c, \qquad n > 1.$$

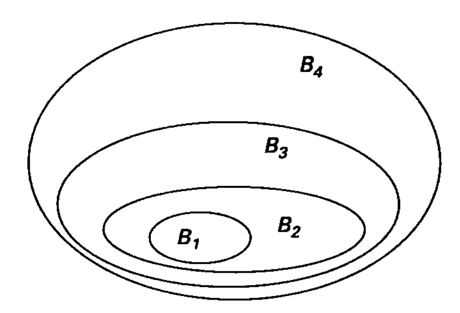


Figure 8.1-9 Illustrating an increasing sequence of events.

The A_n are disjoint and $\bigcup_{n=1}^{N} A_n = \bigcup_{n=1}^{N} B_n$ for all N. Also $B_N = \bigcup_{n=1}^{N} B_n$ because the B_n are increasing. So

$$P[B_N] = P\left[\bigcup_{n=1}^{N} B_n\right] = P\left[\bigcup_{n=1}^{N} A_n\right] = \sum_{n=1}^{N} P[A_n],$$

and

$$\begin{aligned}
\lim_{n\to\infty} P[B_N] &= \lim_{n\to\infty} \sum_{n=1}^{N} P[A_n] \\
&= \sum_{n=1}^{+\infty} P[A_n] \quad \text{by definition of the limit of a sum,} \\
&= P\left[\bigcup_{n=1}^{\infty} A_n\right] \quad \text{by Axiom 4,} \\
&= P[B_\infty] \quad \text{by definition of the } A_n.
\end{aligned}$$

This last step results from $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n \triangleq B_\infty$. ■

Corollary 8.1-1 Let B_n be a *decreasing sequence* of events, that is, $B_n \supset B_{n+1}$ for all $n \geq 1$. Then

$$\lim_{n\to\infty} P[B_n] = P[B_\infty],$$

where

$$B_\infty \triangleq \bigcap_{n=1}^{\infty} B_n.$$

Proof Similar to proof of Theorem 8.1-1 and left to the student. ■

Example 8.1-7

Let $B_n \triangleq \{X[k] < 2 \text{ for } 0 < k < n\}$, for $n = 0, 1, 2, \ldots$. In words, B_n is the event that $X[k]$ is less than 2 for the indicated range of k. Clearly B_{n+1} is a subset of B_n, that is, $B_{n+1} \subset B_n$ for all $n = 0, 1, 2, \ldots$. Also if we set $B_\infty \triangleq \{X[k] < 2 \text{ for all } k \geq 0\}$, then $B_\infty = \cap_{n=1}^{\infty} B_n$. So we can write, by the above corollary,

$$\begin{aligned}
P[B_\infty] &= \lim_{n\to\infty} P[B_n] \\
&= \lim_{n\to\infty} P[X[0] < 2, \ldots, X[n] < 2].
\end{aligned}$$

Thus, the corollary provides a way of calculating events involving an infinite number of random variables by just taking the limit of the probability involving a finite number of

random variables. This type of limiting calculation is often performed in engineering analyses, and typically without explicit justification (i.e., without worrying about the consistency problem mentioned earlier). In this section we have seen that the correctness of the approach rests on a fundamental axiom of probability theory, Axiom 4 (countable additivity).

We next use the continuity of the probability measure P to prove an elementary fact about CDFs.

Example 8.1-8

(*continuity on the right*) The CDF is continuous from the *right*; that is, for $F_X(x) = P[X(\zeta) \leq x]$ [cf. Property (iii) of F_X in Section 2.3], we have

$$\lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right) = F_X(x).$$

To show this, we define

$$B_n \triangleq \left\{\zeta : X(\zeta) \leq x + \frac{1}{n}\right\}$$

and note that B_n is a decreasing sequence of events, where $B_\infty \triangleq \bigcap_{n=1}^{\infty} B_n = \{\zeta : X(\zeta) \leq x\}$ and

$$F_X\left(x + \frac{1}{n}\right) = P[B_n].$$

By application of Corollary 8.1-1, we get

$$\begin{aligned}\lim_{n\to\infty} F_X\left(x + \frac{1}{n}\right) &= \lim_{n\to\infty} P[B_n] = P[B_\infty] \\ &= F_X(x).\end{aligned}$$

Statistical Specification of a Random Sequence

A random sequence $X[n]$ is said to be *statistically specified* by knowing its Nth-order CDFs for all integers $N \geq 1$, and for all times, $n, n+1, \ldots, n+N-1$, that is, if we know

$$\begin{aligned}&F_X(x_n, x_{n+1}, x_{n+2}, \ldots, x_{n+N-1}; n, n+1, \ldots, n+N-1) \\ &\quad \triangleq P[X[n] \leq x_n, X[n+1] \leq x_{n+1}, \ldots, X[n+N-1] \leq x_{n+N-1}],\end{aligned} \tag{8.1-4}$$

where the variables after the semicolon, $n, n+1, \ldots, n+N-1$, indicate the location of the N random variables in this joint CDF. Note that this is an infinite set of CDFs for *each* order N, because we must know the joint CDF at *all* times n, $-\infty < n < +\infty$. Incurring some penalty in notational clarity, we often write the joint CDFs more simply as

$$F_X(x_n, x_{n+1}, \ldots, x_{n+N-1}), \quad \text{for all } n, \text{ and for all } N \geq 1. \tag{8.1-5}$$

We also define Nth-order CDFs for nonconsecutive time parameters,

$$F_X(x_{n_1}, x_{n_2}, \ldots, x_{n_N}; n_1, n_2, \ldots, n_N).$$

It may seem that this statistical specification is some distance from a complete description of the entire random sequence since no one distribution function in this infinite set of finite-order CDFs describes the *entire* random sequence. Nevertheless, if we specify all these finite-order joint distributions at all finite times, using continuity of the probability measure that we have just shown, we can calculate the probabilities of events involving infinite numbers of random variables via limiting operations involving the finite-order CDFs. Of course, we do have to make sure that our set of Nth-order CDFs is consistent within itself! Sometimes it is trivial, for instance, the case where all the random variables that make up the random sequence are independent of one another, for example, a Bernoulli random sequence.

Example 8.1-9

(*consistency*) For consistency, the low-order CDFs must agree with the higher-order CDFs. For example, considering just $N = 2$ and 3, we must have

$$F_X(x_n, x_{n+2}; n, n+2) = F_X(x_n, \infty, x_{n+2}; n, n+1, n+2),$$

for all n, and for all values of x_n and x_{n+2}. Likewise, the $N = 1$ CDFs must be consistent with those of $N = 2$. Further the consistency must extend to all higher orders N.

Consistency can be *guaranteed* by construction, as in the case of the filtered Bernoulli random sequence of Example 8.1-6 above. If we were faced with a suspect set of Nth-order CDFs of unknown origin, it would be a daunting task, indeed, to show that they were consistent. Hence, we see the important role played by constructive models in stochastic sequences and processes.

In summary, we have seen two ways to specify a random sequence: the statistical characterization (Equation 8.1-4) and the direct specification in terms of the random functions $X[n, \zeta]$. We use the word *statistical* to indicate that the former information can be obtained, at least conceptually, by estimating the Nth-order CDFs for $N = 1, 2, 3, \ldots$ and so forth, that is, by using *statistics*.

The Nth-order probability density functions (pdf's) are given for differentiable F_X as

$$f_X(x_n, x_{n+1}, \ldots, x_{n+N-1}; n, n+1, \ldots, n+N-1) = \frac{\partial^N F_X(x_n, x_{n+1}, \ldots, x_{n+N-1}; n, n+1, \ldots, n+N-1)}{\partial x_n \partial x_{n+1} \ldots \partial x_{n+N-1}}, \tag{8.1-6}$$

for every integer (time) n and positive integer (order) N. Sometimes we will omit the subscript X when only one random sequence is under consideration. Also, we may drop the explicit time notation and write

$$f_X(x_n, x_{n+1}, \ldots, x_{n+N-1}) \quad \text{for} \quad f_X(x_n, x_{n+1}, \ldots, x_{n+N-1}; n, n+1, \ldots, n+N-1).$$

We will sometimes want to deal with complex random variables and sequences. By this we mean an ordered pair of real random variables, that is, $X = (X_R, X_I)$ often written as $X = X_R + jX_I$ with CDF

$$F_X(x_R, x_I) \triangleq P[X_R \le x_R, X_I \le x_I].$$

The corresponding pdf is then

$$f_X(x_R, x_I) = \frac{\partial^2 F_X(x_R, x_I)}{\partial x_R \partial x_I}.$$

To simplify notation we will write $f_X(x)$ for $f_X(x_R, x_I)$ in what follows, with the understanding that the respective integrals (sums for discrete valued complex case) are really double integrals on the (x_R, x_I) plane if the random variable is complex.†

The *moments* of a random sequence play an important role in most applications. In part this is because for a large class of random sequences (so-called *ergodic* sequences, to be covered in Section 10.4 in Chapter 10), they can be easy to estimate from just one sample sequence. The first moment or *mean function* of a random sequence is

$$\mu_X[n] \triangleq E\{X[n]\} = \int_{-\infty}^{+\infty} x f_X(x;n)dx$$
$$= \int_{-\infty}^{+\infty} x_n f_X(x_n)dx_n$$

for a continuous-valued random sequence $X[n]$. The mean function for a discrete-valued random sequence, taking on values from the set $\{x_k, -\infty < k < +\infty\}$ at time n, is evaluated as

$$\mu_X[n] = E\{X[n]\} = \sum_{k=-\infty}^{+\infty} x_k P[X[n] = x_k]. \quad (8.1\text{-}7)$$

In the case of a mixed random sequence, as in the case of mixed random variables, it is convenient to write

$$\mu_X[n] = \int_{-\infty}^{+\infty} x f_X(x;n)dx + \sum_{k=-\infty}^{+\infty} x_k P[X[n] = x_k]. \quad (8.1\text{-}8)$$

Actually using the concept of the Stieltjes integral [8-3] both terms can be rewritten in the one form

$$\mu_X[n] = \int_{-\infty}^{+\infty} x\, dF_X(x;n),$$

in terms of the CDF $F_X(x;n)$.

The expected value of the product of the random sequence evaluated at two times $X[k]X^*[l]$ is called the *autocorrelation function* and is a two-parameter function of *both* times k and l, where $-\infty < k, l < +\infty$,

†Complex random sequences are used as equivalent baseband models of certain bandpass signals and noises. The resulting complex valued simulation can be then run at a much lower sample rate.

$$R_{XX}[k,l] \triangleq E\{X[k]X^*[l]\}$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x_k x_l^* f_X(x_k, x_l; k, l) dx_k dx_l, \tag{8.1-9}$$

when the autocorrelation function exists (the usual case, but of course, in some cases the integral might not converge). Most of the time we will deal with *second-order* random sequences, defined by their property of having finite *average power* $E\{|X[n]|^2\} < \infty$. Then the corresponding correlation function will always exist. Later we shall see that the conjugate on the second factor in the autocorrelation function definition results in some notational simplicities for complex-valued random sequences. We will also define the *centered* random sequence $X_c[n] \triangleq X[n] - \mu_X[n]$, which is zero-mean, and consider its autocorrelation function, called the *autocovariance function* of the original sequence $X[n]$. It is defined as

$$K_{XX}[k,l] \triangleq E\{(X[k] - \mu_X[k])(X[l] - \mu_X[l])^*\}. \tag{8.1-10}$$

Directly from these definitions, we note the following symmetry conditions must hold:

$$R_{XX}[k,l] = R^*_{XX}[l,k], \tag{8.1-11}$$

$$K_{XX}[k,l] = K^*_{XX}[l,k], \tag{8.1-12}$$

called *Hermitian symmetry*. Also note that

$$K_{XX}[k,l] = R_{XX}[k,l] - \mu_X[k]\mu^*_X[l]. \tag{8.1-13}$$

The *variance function* is defined as $\sigma^2_X[n] \triangleq K_{XX}[n,n]$ and denotes the average power in $X_c[n]$. The *power* of $X[n]$ itself has been given above and equals $R_{XX}[n,n]$.

Example 8.1-10

(*Example 8.1-1 cont'd.*) The mean function of $X[n]$ as given in Example 8.1-1 is

$$\mu_X[n] = E\{X[n]\} = E\{Xf[n]\} = \mu_X f[n],$$

where μ_X is the mean of the random variable X. The autocorrelation function is

$$R_{XX}[k,l] = E\{X[k]X^*[l]\} = E\{Xf[k]X^*f^*[l]\}$$
$$= E\{|X|^2\}f[k]f^*[l],$$

and so the autocovariance function is given as

$$K_{XX}[k,l] = E\{|X|^2 f[k]f^*[l]\} - |\mu_X|^2 f[k]f^*[l]$$
$$= E\{|X|^2 - |\mu_X|^2\}f[k]f^*[l]$$
$$= E\{|X - \mu_X|^2\}f[k]f^*[l]$$
$$= \sigma^2_X f[k]f^*[l],$$

where $\sigma^2_X = \mathrm{Var}(X)$. We thus see that the variance $\sigma^2_X[n]$ is just $\sigma^2_X|f[n]|^2$.

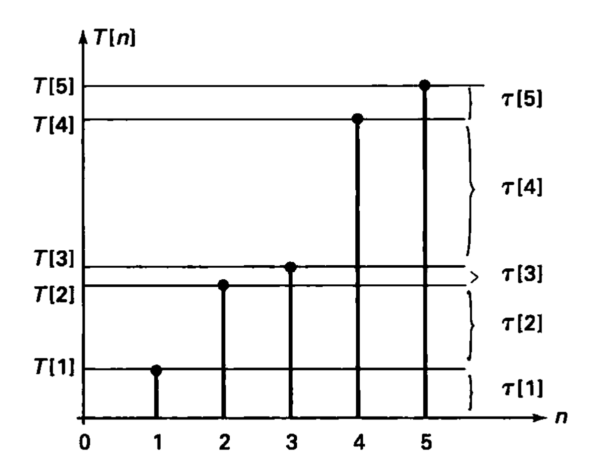


Figure 8.1-10 The $T[n]$ are arrival times and the $\tau[n]$ are interarrival times.

We look at a sequence which fits our notions of randomness better in the next example.

Example 8.1-11

(*waiting times*) Consider the random sequence consisting of i.i.d. random variables $\tau[n]$ for $n \geq 1$, each with the exponential pdf of Equation 2.4-16, that is,†

$$f_T(t;n) = f_T(t) = \lambda \exp(-\lambda t)u(t), \qquad n = 1, 2, \ldots$$

Write the running sum of the $\tau[k]$ up to time n, defined as

$$T[n] \triangleq \sum_{k=1}^{n} \tau[k], \tag{8.1-14}$$

and consider $T[n]$ as a second random sequence for $n = 1, 2, \ldots$. It turns out that the arrival of random events in time is often modeled in this way. We say that $T[n]$ is the *time to the nth arrival* or *waiting time* and we call the $\tau[n]$ the *interarrival times.*‡ See Figure 8.1-10.

Later, in Chapter 9, we shall see that the important Poisson random process can be constructed in this way. Here we want to determine the pdf of $T[n]$ at each n based on the definition in Equation 8.1-14. Using the fact that the $\tau[k]$ are independent, we can apply Equation 4.7-3 and conclude that the pdf of $T[n]$ will be the $(n-1)$-fold convolution product of exponential pdf's. Using convolution to determine the pdf of $T[2]$, we get

$$f_T(t;2) = f_T(t) * f_T(t) = \lambda^2 t \exp(-\lambda t)u(t).$$

†Recall that $\lambda = 1/\mu$.

‡Please regard τ as a "capital tau" to continue our distinction between a random variable and the value it takes on, that is, $X = x$.

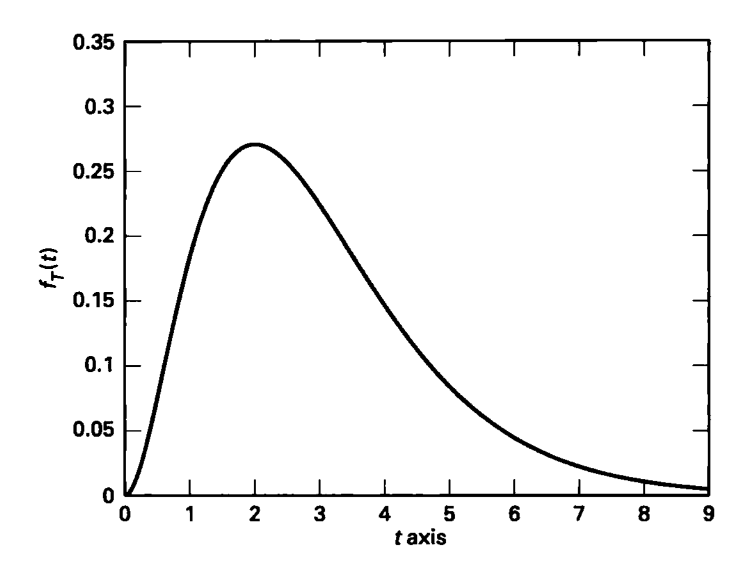


Figure 8.1-11 A plot of the Erlang pdf for $\lambda = 1$ and $n = 3$.

Convolving this result with the exponential pdf a second time, we get

$$f_T(t;3) = \frac{1}{2}\lambda^3 t^2 \exp(-\lambda t)u(t).$$

It turns out that the general form is the Erlang pdf,

$$f_T(t;n) = \frac{(\lambda t)^{n-1}}{(n-1)!}\lambda \exp(-\lambda t)u(t). \tag{8.1-15}$$

The Erlang or gamma pdf [8-4] is widely used in waiting-time problems in telecommunications networks and is plotted via MATLAB in Figure 8.1-11 for $n = 3$ and $\lambda = 1.0$, which is the waiting time for $n = 3$ arrivals.

We can establish this density's correctness by the *Principle of Mathematical Induction.* (See Section A.4 in Appendix A.) It is composed of two steps: (1) First show the formula is correct at $n = 1$; (2) then show that *if* the formula is true at $n-1$, it *must* also be true at n. Combining these two steps, we have effectively proved the result for all positive integers n.

We see that $f_T(t;1)$ in Equation 8.1-15 is correct, so we proceed by assuming Equation 8.1-15 is true at $n-1$. By convolving with the exponential, we can show that it is true at n as follows:

$$\begin{aligned} f_T(t;n) &= f_T(t;n-1) * \lambda \exp(-\lambda t)\, u(t) \\ &= \int_0^t \exp(-\lambda\tau)\frac{(\lambda\tau)^{n-2}}{(n-2)!}\lambda^2 \exp(-\lambda(t-\tau))d\tau\, u(t) \end{aligned}$$

$$= \lambda^n \exp(-\lambda t) \int_0^t \frac{\tau^{n-2}}{(n-2)!} d\tau \, u(t)$$

$$= \lambda^n \exp(-\lambda t) \frac{t^{n-1}}{(n-1)!} u(t).$$

Using the i.i.d. property of the $\tau[n]$, we can also compute the mean as

$$\mu_T[n] = n\mu_\tau = n(1/\lambda) = n/\lambda$$

and variance of the sum $T[n]$ by repeated use of property (A) of Equation 4.3-18.

$$\text{Var}\,[T[n]] = n\text{Var}[\tau] = n/\lambda^2.$$

We next introduce the most widely used random model in electrical engineering, communications, and control: the Gaussian (Normal) random sequence. Its wide popularity stems from two important facts: (1) the Central Limit theorem (Theorem 4.7-2) assures that many processes occurring in practice are approximately Gaussian; and (2) the mathematics is especially tractable in problems involving detection, estimation, filtering, and control theory.

Definition 8.1-3 A random sequence $X[n]$ is called a *Gaussian random sequence* if its Nth-order CDFs (pdf's) are jointly Gaussian, for all $N \geq 1$. ■

We note that the mean and covariance function will specify a Gaussian random sequence in the same way that the mean vector and covariance matrix determine a Gaussian random vector (see Section 5.5). This is because each Nth-order distribution function is just the CDF of a Gaussian random vector whose mean vector and covariance matrix are expressible in terms of the mean and covariance functions of the Gaussian random sequence.

Example 8.1-12

(*pairwise average*) Let $W[n]$ be a real-valued Gaussian i.i.d. sequence with mean $\mu_W[n] = 0$ for all n and autocorrelation function $R_W[k, l] = \sigma^2\delta[k - l]$, $\sigma > 0$, where δ is the *discrete-time impulse*

$$\delta[n] \triangleq \begin{cases} 1, & n = 0, \\ 0, & n \neq 0. \end{cases}$$

If we form a covariance matrix, then, for a vector of any N distinct samples, it will be diagonal. So, by Gaussianity, each Nth-order pdf will factor into a product of N first-order pdf's. Hence the elements of this random sequence are jointly independent, or what we call an *independent (Gaussian) random sequence* (cf. Definition 8.1-2). Next we create the random sequence $X[n]$ by taking the sum of the current and previous $W[n]$ values,

$$X[n] \triangleq W[n] + W[n-1], \quad \text{for } -\infty < n < +\infty.$$

Here $X[n]$ is also Gaussian in all its Nth-order distributions (since a linear transformation of a Gaussian random vector produces a Gaussian vector by Theorem 5.6-1); hence $X[n]$ is also a Gaussian random sequence. We can easily evaluate the mean of $X[n]$ as

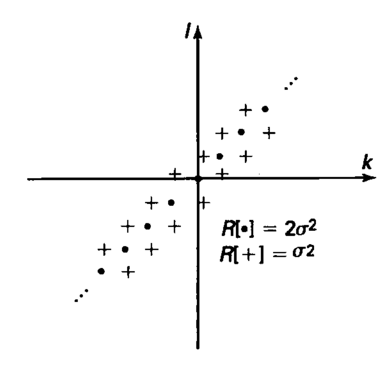


Figure 8.1-12 Diagram of the tri-diagonal correlation function of Example 8.1-12.

$$\mu_X[n] = E\{X[n]\} = E\{W[n]\} + E\{W[n-1]\}$$
$$= 0,$$

and its correlation function as

$$\begin{aligned}
R_{XX}[k,l] &= E\{X[k]X[l]\} \\
&= E\{(W[k] + W[k-1])\,(W[l] + W[l-1])^*\} \\
&= E\{W[k]W[l]\} + E\{W[k]W[l-1]\} \\
&\quad + E\{W[k-1]W[l]\} + E\{W[k-1]W[l-1]\} \\
&= R_{WW}[k,l] + R_{WW}[k,l-1] + R_{WW}[k-1,l] + R_{WW}[k-1,l-1] \\
&= \sigma^2\big(\delta[k-l] + \delta[k-l+1] + \delta[k-l-1] + \delta[k-l]\big).
\end{aligned}$$

We can plot this autocorrelation in the (k,l) plane as shown in Figure 8.1-12 and see the time extent of the dependence of the random sequence $X[n]$.

From this figure, we see that the autocorrelation has value $2\sigma^2$ on the diagonal line $l = k$ and has value σ^2 on the diagonal lines $l = k \pm 1$. It should be clear from Figure 8.1-12 that $X[n]$ is **not** an independent random sequence. However, the banded support of this covariance function signifies that dependence is limited to shifts $(k-l) = \pm 1$ in time. Beyond this lag we have uncorrelated, and hence in this Gaussian case, independent random variables.

Example 8.1-13

(*random walk sequence*) Continuing with infinite-length Bernoulli trials, we now define a random sequence $X[n]$ as the running sum of the number of successes (heads) minus the number of failures (tails) in n trials times a step size s,

$$X[n] = \sum_{k=1}^{n} W[k] \qquad \text{with} \qquad X[0] = 0,$$

where we redefine $W[k] = +s$ for outcome ζ =H and $W[k] = -s$ for outcome ζ =T.

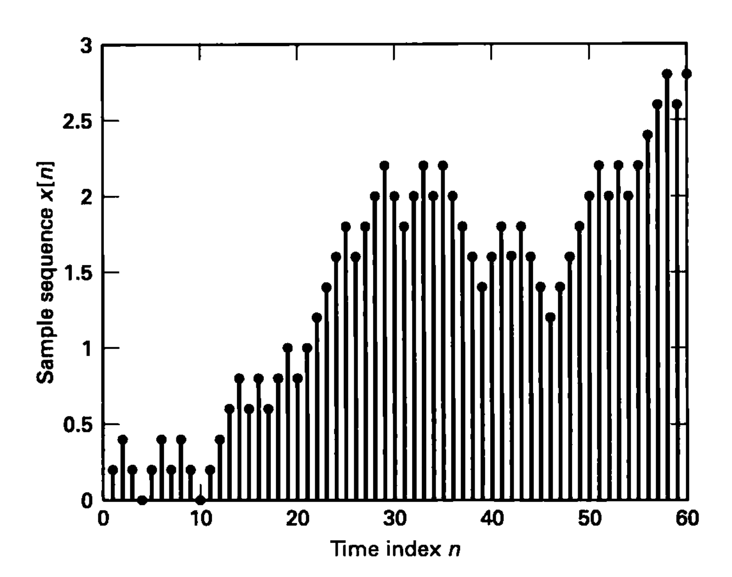


Figure 8.1-13 A sample sequence $x[n]$ for random walk $X[n]$ with step size $s = 0.2$.

The resulting sequence then models a *random walk* on the integers starting at position $X[0] = 0$. At each succeeding time unit a step of size s is taken either to the right or to the left. After n steps we will be at a position rs for some integer r. This is illustrated in Figure 8.1-13.

If there are k successes and necessarily $(n - k)$ failures, then we have the following relation:

$$\begin{aligned} rs &= ks - (n-k)s \\ &= (2k - n)s, \end{aligned}$$

which implies that $k = (n+r)/2$, for those values of r that make the right-hand side an integer. Then with $P[\text{success}] = P[\text{failure}] = \frac{1}{2}$, we have

$$\begin{aligned} P\{X[n] = rs\} &= P\left[(n+r)/2 \text{ successes}\right] \\ &= \begin{cases} \binom{n}{(n+r)/2} 2^{-n}, & (n+r)/2 \text{ an integer, } r \le n \\ 0, & \text{else.} \end{cases} \end{aligned}$$

Using the fact that $X[n] = W[1] + W[2] + \ldots + W[n]$ and that the W's are jointly independent, we can compute the mean and variance of the random walk as follows:

$$E\{X[n]\} = \sum_{k=1}^{n} E\{W[k]\} = \sum_{k=1}^{n} 0 = 0,$$

and

$$\begin{aligned}E\{X^2[n]\} &= \sum_{k=1}^{n} E[W^2[k]] \\ &= \sum_{k=1}^{n} 0.5[(+s)^2 + (-s)^2] \\ &= ns^2.\end{aligned}$$

If we normalize $X[n]$ by dividing $\sqrt{n}$ and define

$$\tilde{X}[n] \triangleq \frac{1}{\sqrt{n}} X[n],$$

then by the Central Limit Theorem 4.7-2 we have that the CDF of $\tilde{X}[n]$ converges to the Gaussian (Normal) distribution $N(0, s^2)$. Thus for n large enough, we can approximate the probabilities

$$P[a < \tilde{X}[n] \leq b] = P[a\sqrt{n} < X[n] \leq b\sqrt{n}] \simeq \text{erf}(b/s) - \text{erf}(a/s).$$

Note, however, that when this probability is small, very large values of n might be required to keep the percentage error small because small errors in the CDF may be comparable to the required probability value. In practice this means that the Normal approximation will not be dependable on the tails of the distribution but only in the central part, hence the name Central Limit Theorem.

Note also that while $X[n]$ can never be considered approximately Gaussian for any n (e.g., if n is even, $X[n]$ can only be an even multiple of s), still we can approximately calculate the probability

$$\begin{aligned}P[(r-2)s < X[n] \leq rs] &= P\left[\frac{(r-2)s}{\sqrt{n}} < \tilde{X}[n] \leq \frac{rs}{\sqrt{n}}\right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{(r-2)/\sqrt{n}}^{r/\sqrt{n}} \exp(-0.5v^2)dv \\ &\approx 1/\sqrt{\pi(n/2)} \exp(-r^2/2n),\end{aligned}$$

where r is small with respect to $\sqrt{n}$. See Section 1.11 for a similar result. In obtaining the last line, we assumed that the integrand was approximately constant over the interval $[(r-2)/\sqrt{n}, r/\sqrt{n}]$.

The waiting-time sequence in Example 8.1-11 and the random walk in Example 8.1-13 both have the property that they build up over time from independent components or increments. More generally we can define an independent-increments property.

Definition 8.1-4 A random sequence is said to have *independent increments* if for all integer parameters $n_1 < n_2 < \ldots < n_N$, the increments $X[n_1], X[n_2] - X[n_1], X[n_3] - X[n_2], \ldots, X[n_N] - X[n_{N-1}]$ are jointly independent for all integers $N > 1$. ■

If a random sequence has independent increments, one can build up its Nth-order probabilities (PMFs and pdf's) as products of the probabilities of its increments. (See Problem 8.10.)

In contrast to the evolving nature of independent increments, many random sequences have constant statistical properties that are invariant with respect to the index parameter n, normally time or distance. When this is valid, the random model is simplified in two ways: First, it is time-invariant, and second, the usually small number of model parameters can be estimated from available data.

Definition 8.1-5 If for all orders N and for all shift parameters k, the joint CDFs of $(X[n], X[n+1], \ldots, X[n+N-1])$ and $(X[n+k], X[n+k+1], \ldots, X[n+k+N-1])$ are the same functions, then the random sequence is said to be *stationary*, i.e., for all $N \geq 1$,

$$
\begin{aligned}
&F_X(x_n, x_{n+1}, \ldots, x_{n+N-1}; n, n+1, \ldots, n+N-1) \\
&\quad = F_X(x_n, x_{n+1}, \ldots, x_{n+N-1}; n+k, n+1+k, \ldots, n+N-1+k) \qquad (8.1\text{-}16)
\end{aligned}
$$

for all $-\infty < k < +\infty$ and for all x_n through x_{n+N-1}. This definition also holds for pdf's when they exist and PMFs in the discrete amplitude case. ■

If we look back at Example 8.1-12, we see that $X[n]$ and $W[n]$ are both stationary random sequences. The same was true of the interarrival times $\tau[n]$ in Example 8.1-11, but the random arrival or waiting time sequence $T[n]$ was clearly nonstationary, since its mean and variance increase with time n.

Note that stationarity does not mean that the sample sequences all look "similar," or even that they all look "noisy."† Also, unlike the concept of stationarity in mathematics and physics, we don't directly characterize the realizations of the random sequence as stationary, just the deterministic functions that characterize their behavior, i.e., CDF, PMF, and pdf.

It is often desirable to partially characterize a random sequence based on knowledge of only its first two moments, that is, its mean function and covariance function. This has already been encountered for random vectors in Chapter 5. We will encounter this for random sequences when we present a discussion of linear estimation in the signal-processing applications of Chapter 11. In anticipation we define a weakened kind of stationarity that involves only the mean and covariance (or correlation) functions. Specifically, if these two functions are consistent with stationarity, then we say that the random sequence is *wide-sense stationary* (WSS).

†For example, suppose we do the Bernouilli experiment of flipping a fair coin once and generate a random sequence as follows: If the outcome is *heads* then $X[n] = 1$ for all n. If the outcome is tails then $X[n] = W[n]$, that is, stationary white noise again for all n. Thus, the sample sequences look quite dissimilar, but the random sequence is easily seen to be stationary. In Chapter 10, we discuss the property of *ergodicity*, which, loosely speaking, enables expectations (ensemble averages) to be computed from time averages. In this case the sample functions would tend to have the same features; that is, a viewer would subjectively feel that they come from the same source.

Definition 8.1-6 A random sequence $X[n]$ defined for $-\infty < n < +\infty$ is called *wide-sense stationary* (WSS) if

(1) The mean function of $X[n]$ is constant for all integers n, $-\infty < n < +\infty$,

$$\mu_X[n] = \mu_X[0] \quad \text{and}$$

(2) For all times k, l, $-\infty < k, l < +\infty$, and integers n, $-\infty < n < +\infty$, the covariance (correlation) function is independent of the shift n,

$$K_{XX}[k, l] = K_{XX}[k + n, l + n]. \quad \blacksquare \tag{8.1-17}$$

We will call such a covariance (correlation) function *shift-invariant.* If we think of $[k, l]$ as a constellation or set of two samples on the time line, then we are translating this constellation up and down the time line, and saying that the covariance function does not change. When the mean function is constant, then shift invariance of the covariance and correlation functions is equivalent. Otherwise it is not. For a constant mean function, we can check property (2) for either the covariance or correlation function.

While all stationary sequences are WSS, the reverse is not true. For example, the third moment could be shift-variant in a manner not consistent with stationarity even though the first moment is constant and the second moment is shift-invariant. Then the random sequence would be WSS but not stationary. To further distinguish them, sometimes we refer to stationarity as *strict-sense stationarity* to avoid confusion with the weaker concept of wide-sense stationarity.

Theorem 8.1-2 All stationary random sequences are WSS.

Proof We first show that the mean is constant for a stationary random sequence. Let n be arbitrary

$$\mu_X[n] = E\{X[n]\} = \int_{-\infty}^{+\infty} x f_X(x; n) dx = \int_{-\infty}^{+\infty} x\, f_X(x; 0) dx = \mu_X[0],$$

since $f_X(x; n)$ does not depend on n. Next we show that the covariance function is shift-invariant by first showing that the correlation is shift-invariant:

$$\begin{aligned} R_{XX}[k, l] &= E\{X[k]X^*[l]\} \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_k x_l^* f_X(x_k, x_l) dx_k dx_l \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_{n+k} x_{n+l}^* f_X(x_{n+k}, x_{n+l}) dx_{n+k} dx_{n+l},^{\dagger} \\ &= R_{XX}[n + k, n + l], \end{aligned}$$

†These middle two lines use our simplified notation. They are not trivially equal because $f_X(x_k, x_l)$ and $f_X(x_{k+n}, x_{l+n})$ are really the joint densities at two different pairs of times. This can be made clear using the full notation: $f_X(x_k, x_l; k, l)$.

since $f_X(x_k, x_l)$ doesn't depend on the shift n, and the x_i's are dummy variables. Finally, we use Equation 8.1-13 and the result on the mean functions to conclude that the covariance function is also shift-invariant. Since the covariance function is shift-invariant for any WSS random sequence, we can define a one-parameter covariance function to simplify the notation for WSS sequences

$$K_{XX}[m] \triangleq E\{X_c[k+m]X_c^*[k]\} = K_{XX}[k+m,k]$$
$$= K_{XX}[m,0]. \qquad (8.1\text{-}18)$$

We also do the same for correlation functions. Writing the one-parameter correlation function in terms of the corresponding two-parameter correlation function, we have

$$R_{XX}[m] = R_{XX}[k+m,k] = R_{XX}[m,0]. \quad \blacksquare$$

Example 8.1-14

(*WSS covariance function*) The covariance function of Example 8.1-12 is shift-invariant and so we can take advantage of the simplified notation. We can thus write $K_{XX}[m] = \sigma^2(2\delta[m] + \delta[m-1] + \delta[m+1])$.

Example 8.1-15

(*two-state random sequence with memory*) We construct the two-level (binary) random sequence $X[n]$ on $n \geq 0$ as follows. Recursively, and for each n (>0) in succession, and for each level, we set $X[n] = X[n-1]$ with probability p, for some given $0 < p < 1$. Otherwise, and with probability $q \triangleq 1-p$, we set $X[n]$ to the "other" value (level). Let the two levels be denoted a and b, and start off the sequence with $X[0] = a$. When $p = 0.5$, this is a special case of the Bernoulli random sequence. When $p \neq 0.5$, this is not an independent random sequence, since $P_X(x_n|x_{n-1}; n, n-1) \neq P_X(x_n; n)$. We say the random sequence has *memory*. To see this, consider the case where $p \approx 1.0$; then set x_n to the level other than x_{n-1}, and note that the conditional transition probability $P_X(x_n|x_{n-1}; n, n-1) \approx 0$, while the unconditional probability $P_X(x_n; n)$ is not so constrained. In fact, $P_X(e_n; n)$ would not be expected to favor either level, since the above transition rules are the same for either level. Intuitively, at least, it makes sense to call $X[n-1]$ the *state* at time $n-1$. In fact, the rules for generating this random sequence can be summarized in the *state-transition diagram* shown in Figure 8.1-14, where the directed branches are labeled by the relevant probabilities for the next state, given the present state, as can easily be verified by inspection. We can refer to p as the *no-transition* probability. This is a first example of a Markov random sequence which will be studied in Section 8.5.

The following MATLAB m-file can generate sample functions for these random sequences on $n \geq 1$:

```
function[w]=randmemseq(p,N,w0,a,b)
w=a*ones(1,N);
w(1)=w0;
for i=2:N
    rnum=rand;
    if rnum <p;
```

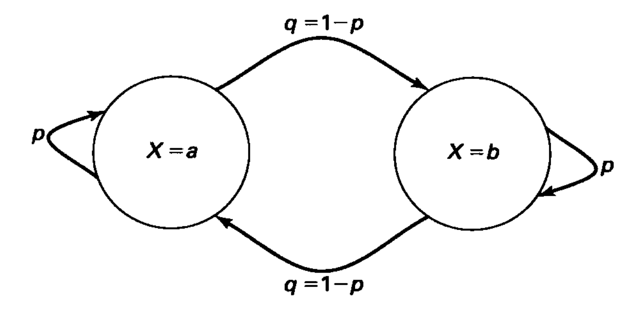


Figure 8.1-14 State-transition diagram of two-state (binary) random sequence with memory.

```
     w(i)=w(i-1);
     else
    if w(i-1)==a;
       w(i)=b;
    else
       w(i)=a;
    end
end
stem(1:N,w)
title('random sequence with memory')
xlabel('discrete time')
ylabel('level')
end
```

Sample waveforms are given in Figures 8.1-15 to 8.1-17 corresponding to level values $b = 1$, $a = 0$, and several values of p. We note that when p is near 1, there are few transitions. For p near 0.5, there will be many transitions displaying little memory. When $p = 0$, there is a transition every time.

Example 8.1-16

(*correlation function of random sequence with memory*) Assume that the random sequence with memory of the last example has been running for a very long time. Later on we will show that in this case, a steady state develops wherein the probabilities of the two levels are constant with time and independent of the starting state (level). Here we assume that the steady state holds for all finite time. Clearly from the symmetry shown in the state diagram, it must be that $P_X(a) = P_X(b) = 0.5$. Now assume that the lower level $a = 0$ and the upper level is b as before, and consider the correlation at two distinct times n and $n + k$. We can write

$$
\begin{aligned}
R_{XX}[n, n+k] &= b^2 P_X(b, b; n, n+k) \\
&= b^2 P(X[n] = b) P(X[n+k] = b | X[n] = b) \\
&= (b^2/2)\, P(X[n+k] = b | X[n] = b),
\end{aligned}
$$

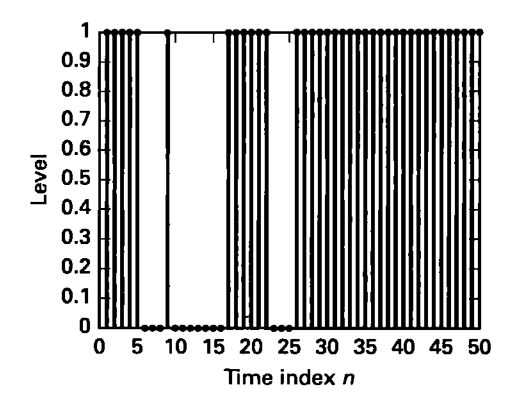


Figure 8.1-15 Initial level $X[1] = 1$, no-transition probability $p = 0.8$.

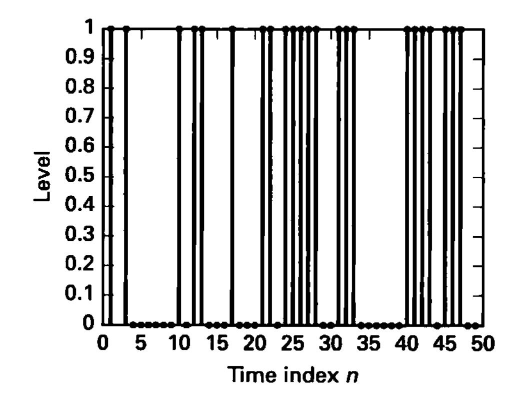


Figure 8.1-16 Initial level $X[1] = 1$, no-transition probability $p = 0.5$, the Bernoulli case.

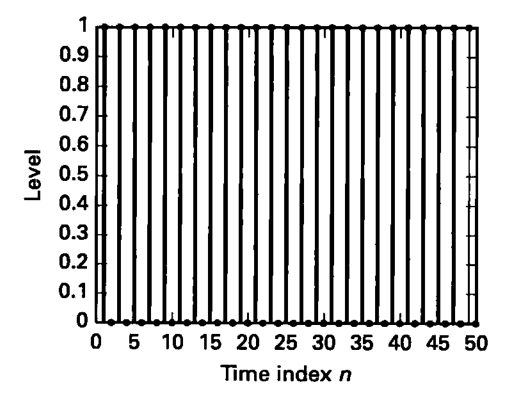


Figure 8.1-17 Initial value $X[1] = 1$, no-transition probability $p = 0$.

where the first equality holds since all the terms involving a are zero since $a = 0$. Now the only way that X can equal b at both times n and $n+k$ is for an even number of transitions to occur between these two times, and the probability of this is given by

$$P\{\text{even number of transitions}\} = \sum_{l=0,2,4,\ldots}^{k} \binom{k}{l} (1-p)^l p^{k-l}$$
$$\triangleq A_e,$$

which follows from the fact that this is just Bernoulli trials with "success" = "transition" and "failure" = "no transition." Thus interchanging the usual role of p and q in Bernoulli trials, we just add up the probability of an even number of successes (transitions). It turns out that A_e can be evaluated in closed form by the following "trick." Define

$$A_o \triangleq \sum_{l=1,3,5,\ldots}^{k} \binom{k}{l} (1-p)^l p^{k-l}(-1)^l.$$

Clearly, we have $A_e - A_o = 1$ since l is always odd valued in the sum A_o. Similarly we note that

$$A_e = \sum_{l=0,2,4,\ldots}^{k} \binom{k}{l} (1-p)^l p^{k-l}(-1)^l$$
$$= \sum_{l=0,2,4,\ldots}^{k} \binom{k}{l} (p-1)^l p^{k-l},$$

where the first equality holds because l is always even in A_e. We now can see that

$$A_e + A_o = \sum_{l=0}^{k} \binom{k}{l} (p-1)^l p^{k-l}$$
$$= (2p-1)^k,$$

by the Binomial Theorem. It follows at once that $A_e = (1/2)\left[(2p-1)^k + 1\right]$, so that

$$R_{XX}[n, n+k] = (b^2/4)\left[(2p-1)^k + 1\right],$$

which shows that $X[n]$ is WSS. We can write this correlation function more cleanly for the case $p > 1/2$. On defining $\alpha \triangleq \ln(2p-1)|$, we have

$$R_{XX}[k] = (b^2/4)\left[\exp(-\alpha\,|k|) + 1\right].$$

Also since the mean value of $X[n]$ is easily seen to be $b/2$, we get the autocovariance function

$$K_{XX}[k] = (b^2/4)\exp(-\alpha\,|k|).$$

A MATLAB m-file for displaying the covariance functions of these sequences, for three values of p, is shown below:

```
function[mc1,mc2,mc3]=markov(b,p1,p2,p3,N)
mc1=0*ones(1,N);
mc2=0*ones(1,N);
mc3=0*ones(1,N);
for i=1:N
   mc1(i)=0.25*(b^2)*(((2*p1-1)^(i-1)));
   mc2(i)=0.25*(b^2)*(((2*p2-1)^(i-1)));
   mc3(i)=0.25*(b^2)*(((2*p3-1)^(i-1)));
end
x=linspace(0,N-1,N);
plot(x,mc1,x,mc2,x,mc3)
title('covariance of Markov Sequences')
xlabel('Lag interval')
ylabel('covariance value')
```

The normalized covariances for $p = 0.8$, 0.5, and 0.2 and $b = 2$ are shown in Figure 8.1-18.

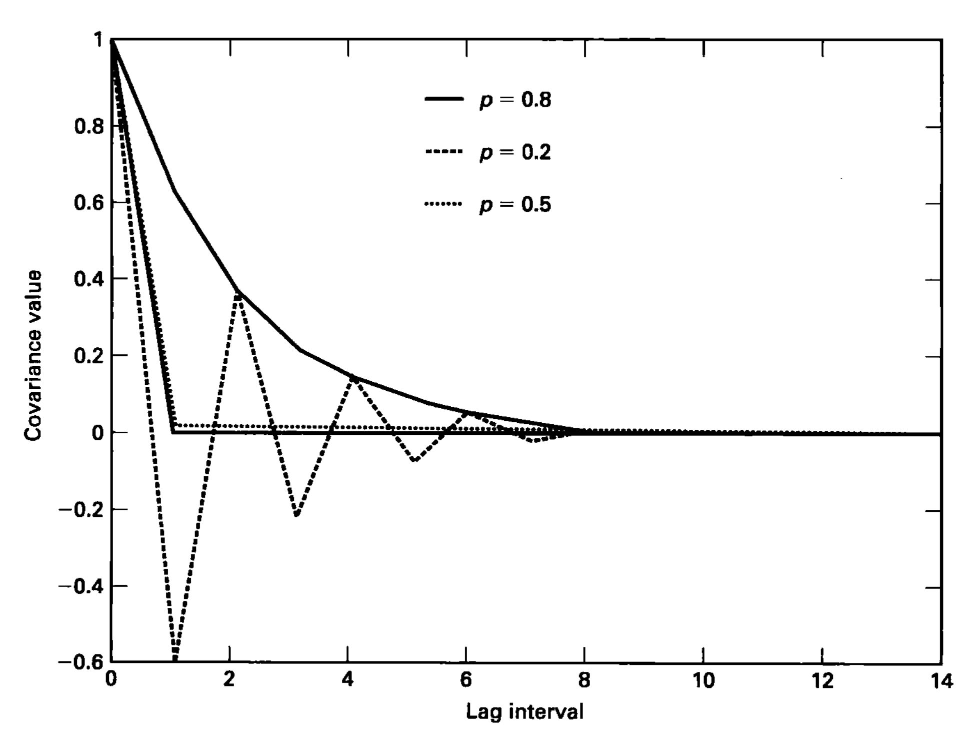


Figure 8.1-18 The covariance functions for different values of the parameter p. (Points connected by straight lines.)

This type of random sequence, which exhibits a one-step memory, is called a *Markov* random sequence (there are variations on the spelling of Markov) in honor of the mathematician A. A. Markov (1856–1922). In Section 8.5 we shall discuss this class of random sequences in greater detail. In the meanwhile we note that the system discussed in Example 8.1-5, that is, $X[n] = \alpha X[n-1] + W[n]$, also exhibited a one-step memory and, hence could also be regarded as a Markov sequence, when $W[n]$ is an independent random sequence.

In Section 8.2, we provide a review or summary of the theory of linear systems for sequences, that is, discrete-time linear system theory. Readers with adequate background may skip this section. In Section 8.3, we will apply this theory to study the effect of linear systems on random sequences, an area rich in applications in communications, signal processing, and control systems.

8.2 BASIC PRINCIPLES OF DISCRETE-TIME LINEAR SYSTEMS

In this section we present some fundamental material on discrete-time linear system theory. This will then be extended in the next section to the case of random sequence inputs and outputs. This material is very similar to the continuous-time linear system theory including the topics of differential equations, Fourier transforms, and Laplace transforms. The corresponding quantities in the discrete-time theory are difference equations, Fourier transforms (for discrete-time signals), and Z-transforms.

With reference to Figure 8.2-1 we see that a linear system can be thought of as having an infinite-length sequence $x[n]$ as input with a corresponding infinite-length sequence $y[n]$ as output. Representing this linear operation in equation form we have

$$y[n] = L\{x[n]\}, \tag{8.2-1}$$

where the linear operator L is defined to satisfy the following definition adapted to the case of discrete-time signals. This notation might appear to indicate that $x[n]$ at time n is the only input value that affects the output $y[n]$ at time n. In fact, all input values can potentially affect the output at any time n. This is why we call L an operator[†] and not merely a function. The examples below will make this point clear. Mathematicians

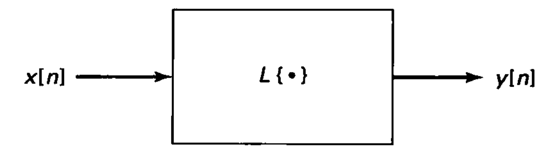


Figure 8.2-1 System diagram for generic linear system $L\{\cdot\}$ with input $x[n]$ and output $y[n]$ and time index parameter n.

[†]Operators map functions (sequences) into functions (sequences).

use the operator notation $y = L\{x\}$ which avoids this difficulty but makes the functional dependence of x and y on the (time) parameter n less clear than in our engineering notation.

Definition 8.2-1 We say a system with operator L is *linear* if for all permissible input sequences $x_1[n]$ and $x_2[n]$, and for all permissible pairs of scalar gains a_1 and a_2, we have

$$L\{a_1x_1[n] + a_2x_2[n]\} = a_1L\{x_1[n]\} + a_2L\{x_2[n]\}. \quad \blacksquare$$

In words, the response of a linear system to a weighted sum of inputs is the weighted sum of the individual outputs. Examples of linear systems would include *moving averages* such as

$$y[n] = 0.33(x[n+1] + x[n] + x[n-1]), \qquad -\infty < n < +\infty,$$

and *autoregressions* such as,

$$y[n] = ay[n-1] + by[n-2] + cx[n], \qquad 0 \leq n < +\infty,$$

when the initial conditions are *zero*. Both these equations are special cases of the more general *linear constant-coefficient difference equation* (LCCDE),

$$y[n] = \sum_{k=1}^{M} a_k y[n-k] + \sum_{k=0}^{N} b_k x[n-k]. \tag{8.2-2}$$

Example 8.2-1

(*solution of difference equations*) Consider the following second-order LCCDE,

$$y[n] = 1.7y[n-1] - 0.72y[n-2] + u[n], \tag{8.2-3}$$

with $y[-1] = y[-2] = 0$ and $u[n]$ the unit-step function. To solve this equation for $n \geq 0$, we first find the general solution to the *homogeneous equation*

$$y_h[n] = 1.7y_h[n-1] - 0.72y_h[n-2].$$

We try $y_h[n] = Ar^n$, where A and r are to be determined,† and obtain

$$A(r^n - 1.7r^{n-1} + 0.72r^{n-2}) = 0$$

†A thorough treatment of the solution of linear difference equations may be found in [8-5].

or

$$Ar^{n-2}(r^2 - 1.7r + 0.72) = 0.$$

We thus see that any value of r satisfying the *characteristic equation*

$$r^2 - 1.7r + 0.72 = 0$$

will give a general solution to the homogeneous equation. In this case there are two roots at $r_1 = 0.8$ and $r_2 = 0.9$. By linear superposition the general *homogeneous solution* must be of the form†

$$y_h[n] = A_1 r_1^n + A_2 r_2^n,$$

where the constants A_1 and A_2 may be determined from the initial conditions.

To obtain the *particular solution*, we first observe that the input sequence $u[n]$ equals 1 for $n \geq 0$. Thus we try as a particular solution a constant, that is, following standard practice,

$$y_p[n] = B \qquad \text{for } n \geq 0$$

and obtain

$$B - 1.7B + 0.72B = 1$$

or

$$B = 1/(1 - 1.7 + 0.72) = 1/(0.02) = 50.$$

More generally this method can be modified for any input function of the form $C\rho^n$ over adjoining time intervals $[n_1, n_2 - 1]$. One just assumes the corresponding form for the solution and determines the constant C as shown. In this approach, we would solve the difference equation for each time interval separately, piecing the solution together at the boundaries by carrying across final conditions to become the initial conditions for the next interval. We illustrate our approach here for the time interval starting at $n = 0$. The total solution is

$$\begin{aligned} y[n] &= y_h[n] + y_p[n] \\ &= A_1(0.8)^n + A_2(0.9)^n + 50 \qquad \text{for } n \geq 0. \end{aligned}$$

To determine A_1 and A_2, we first evaluate Equation 8.2-3 at $n = 0$ and $n = 1$ using $y[-1] = y[-2] = 0$ to carry across the initial conditions to obtain $y[0] = 1$ and $y[1] = 2.7$, from which we obtain the linear equations

$$A_1 + A_2 + 50 = 1 \quad (\text{at } n = 0)$$

†Since the two roots are less than one in magnitude, the solution will be *stable* when run forward in time index n (cf. [8-5]).

and

$$A_1(0.8) + A_2(0.9) + 50 = 2.7 \quad (\text{at } n = 1).$$

This can be put in matrix form

$$\begin{bmatrix} 1.0 & 1.0 \\ 0.8 & 0.9 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} -49.0 \\ -47.3 \end{bmatrix}$$

and solved to yield

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 32 \\ -81 \end{bmatrix}.$$

Thus the complete solution, valid for $n \geq 0$, is

$$y[n] = 32(0.8)^n - 81(0.9)^n + 50.$$

We could then write the solution for all time, if the system was at rest for $n < 0$, as

$$y[n] = \{32(0.8)^n - 81(0.9)^n + 50\}\, u[n].$$

Note that the LCCDE in the previous example is a linear system because the initial conditions, that is, $y[-1], y[-2]$, were zero, often called the *initial rest condition.* Without initial rest, an LCCDE is not a linear system. More generally, linear systems are described by *superposition* with a possibly *time-variant impulse response*

$$h[n, k] \triangleq L\{\delta[n - k]\}.$$

In words we call $h[n, k]$ the response at time n to an impulse applied at time k. We derive the result by simply writing the input as $x[n] = \sum x[k]\delta[n - k]$, and then using linearity to conclude

$$\begin{aligned} y[n] &= L\left\{ \sum_{k=-\infty}^{+\infty} x[k]\delta[n-k] \right\} \\ &= \sum_{k=-\infty}^{+\infty} x[k] L\{\delta[n-k]\} \\ &= \sum_{k=-\infty}^{+\infty} x[k]\, h[n, k], \end{aligned}$$

which is called the *superposition summation* representation for linear systems.

Many linear systems are made of constant components and have an effect on input signals that is invariant to when the signal arrives at the system. A linear system is called *linear time-invariant* (LTI) or, equivalently, *linear shift-invariant* (LSI) if the response to

a delayed (shifted) input is just the delayed (shifted) response. More precisely, we have the following.

Definition 8.2-2 A linear system L is called *shift-invariant* if for all integer shifts k, $-\infty < k < +\infty$, we have

$$y[n+k] = L\{x[n+k]\} \quad \text{for all } n. \quad \blacksquare \tag{8.2-4}$$

An important property of LSI systems is that they are described by *convolution*,† that is, L is a convolution operator,

$$y[n] = h[n] * x[n] = x[n] * h[n],$$

where

$$h[n] * x[n] \triangleq \sum_{k=-\infty}^{+\infty} h[k]x[n-k], \tag{8.2-5}$$

and the sequence

$$h[n] \triangleq L\{\delta[n]\},$$

is called the *impulse response*. With relation to the time-varying impulse response $h[n,k]$, we can see that $h[n] = h[n,0]$ when a linear system is shift-invariant.

In words we can say that—just as for continuous-time systems—if we know the impulse response of an LSI system, then we can compute the response to any other input by carrying out the convolution operation. In the discrete-time case this convolution operation is a summation rather than an integration, but the operation is otherwise the same.

While in principle we could determine the output to any input, given knowledge of the impulse response, in practice the calculation of the convolution operation may be tedious and time consuming. To facilitate such calculations and also to gain added insight, we turn to a frequency-domain characterization of LSI systems. We begin by defining the *Fourier transform* (*FT*) for sequences as follows.

Definition 8.2-3 The *Fourier transform* for a discrete-time signal or sequence is defined by the infinite sum (if it exists)

$$X(\omega) = FT\{x[n]\} \triangleq \sum_{n=-\infty}^{+\infty} x[n]e^{-j\omega n}, \quad \text{for } -\pi \le \omega \le +\pi,$$

and the function $X(\omega)$ is periodic with period 2π outside this range. The *inverse Fourier transform* is given as

$$x[n] = IFT\{X(\omega)\} = \frac{1}{2\pi}\int_{-\pi}^{+\pi} X(\omega)e^{j\omega n}d\omega. \quad \blacksquare$$

†We encountered the operation of convolution in Chapter 3 when we computed the pdf of the sum of two independent RVs.

One can see that the Fourier transform and its inverse for sequences are really just the familiar Fourier series with the sequence x playing the role of the Fourier coefficients and the Fourier transform X playing the role of the periodic function. Thus, the existence and uniqueness theorems of Fourier series are immediately applicable here to the Fourier transform for discrete-time signals. Note that the frequency variable ω is sometimes called *normalized frequency* because, if the sequence $x[n]$ arose from sampling, the period of such sampling has been lost. It is as though the sample period were $T = 1$, as would be consistent with the $[-\pi, +\pi]$ frequency range of the Fourier transform $X(\omega)$.†

For an LSI system the Fourier transform is particularly significant owing to the fact that complex exponentials are the *eigenfunctions* of discrete-time linear systems, that is,

$$L\{e^{j\omega n}\} = H(\omega)e^{j\omega n}, \tag{8.2-6}$$

as long as the impulse response h is absolutely summable. For LSI systems this absolute summability can easily be seen to be equivalent to bounded-input bounded-output (BIBO) stability [8-5].

Just as in continuous-time system theory, multiplication of Fourier transforms corresponds to convolution in the time (or space) domain.

Theorem 8.2-1 (*convolution theorem*) The convolution,

$$y[n] = x[n] * h[n], \qquad -\infty < n < +\infty,$$

is equivalent in the transform domain to

$$Y(\omega) = X(\omega)H(\omega), \qquad -\pi \le \omega \le +\pi.$$

Proof

$$\begin{aligned}
Y(\omega) &= \sum_{-\infty}^{+\infty} y[n]e^{j\omega n} = \sum_{-\infty}^{+\infty} (x[n] * h[n])\, e^{-j\omega n} \\
&= \sum_n \sum_k x[k]h[n-k]e^{-j\omega n} = \sum_n \sum_k x[k]h[n-k]e^{-j\omega(n-k+k)} \\
&= \sum_n \sum_k [x[k]e^{-j\omega k}h[n-k]e^{-j\omega(n-k)}] \\
&= \sum_k x[k]e^{-j\omega k}\left(\sum_n h[n-k]e^{-j\omega(n-k)}\right) \\
&= \sum_k x[k]e^{-j\omega k}H(\omega) \\
&= X(\omega)H(\omega).
\end{aligned}$$

†If the sequence arose from sampling with sample period T, the (true) radian frequency $\mathbf{\Omega} = \omega/T$.

Thus, discrete-time linear shift-invariant systems are easily understood in the frequency domain similar to the situation for continuous-time LSI systems. Analogous to the Laplace transform for continuous-time signals, there is the Z-transform for discrete-time signals. It is defined as follows. ■

Definition 8.2-4 The Z-transform of a discrete-time signal or sequence is defined as the infinite summation (if it exists)

$$\mathsf{X}(z) \triangleq \sum_{n=-\infty}^{+\infty} x[n]z^{-n}, \tag{8.2-7}$$

where z is a complex variable in the *region of absolute convergence* of this infinite sum.[†] ■

Note that $\mathsf{X}(z)$ is a function of a *complex variable*, while $X(\omega)$ is a function of a *real variable*. The two are related by $\mathsf{X}(z)|_{z=e^{j\omega}} = X(\omega)$. We thus see that, if the Z-transform exists, the Fourier transform is just the restriction of the Z-transform to the unit circle in the complex z-plane. Similarly to the proof of Theorem 8.2-1, it is easy to show that the convolution-multiplication property Equation 8.2-1 is also true for Z-transforms. Analogous to continuous-time theory, the Z-transform $\mathsf{H}(z)$ of the impulse response $h[n]$ of an LSI system is called the *system function*. For more information on discrete-time signals and systems, the reader is referred to [8-5].

8.3 RANDOM SEQUENCES AND LINEAR SYSTEMS

In this section we look at the topic of linear systems with random sequence inputs. In particular we will look at how the mean and covariance functions are transformed by both linear and LSI systems. We will do this first for the general case of a nonstationary random sequence and then specialize to the more common case of a stationary sequence. The topics of this section are perhaps the most widely used concepts from the theory of random sequences. Applications arise in communications when analyzing signals and noise in linear filters, in digital signal processing for the analysis of quantization noise in digital filters, and in control theory to find the effect of disturbance inputs on an otherwise deterministic control system.

The first issue is the meaning of inputing a random sequence to a linear system. The problem is that a random sequence is not just one sequence but a whole family of sequences indexed by the parameter ζ, a point (outcome) in the sample space. As such for each fixed ζ, the random sequence is just an ordinary sequence that may be a permissible input for the linear system. Thus, when we talk about a linear system with a random sequence input, it is natural to say that for each point in the sample space Ω, we input the corresponding realization, that is, the sample sequence $x[n]$. We would therefore regard the corresponding output $y[n]$ as a sample sequence[‡] corresponding to the same point ζ in the sample space, thus collectively defining the output random sequence $Y[n]$.

[†]Note the sans serif font to distinguish between the Z-transform and the Fourier transform.
[‡]Recall that $x[n]$, $y[n]$ denote $X[n,\zeta]$, $Y[n,\zeta]$, respectively, for *fixed* ζ.

Definition 8.3-1 When we write $Y[n] = L\{X[n]\}$ for a random sequence $X[n]$ and a linear system L, we mean that for each $\zeta \in \Omega$ we have

$$Y[n, \zeta] = L\{X[n, \zeta]\}.$$

Equivalently, for each sample function $x[n]$ taken on by the input random sequence $X[n]$, we set $y[n]$ as the corresponding sample sequence of the output random sequence $Y[n]$, that is, $y[n] = L\{x[n]\}$. ■

This is the simplest way to treat systems with random inputs. A difficulty arises when the input sample sequences do not "behave well," in which case it may not be possible to define the system operation for every one of them. In Chapter 10 we will generalize this definition and discuss a so-called mean-square description of the system operation, which avoids such problems, although of necessity it will be more abstract.

In most cases it is very hard to find the probability distribution of the output from the probabilistic description of the input to a linear system. The reason is that since the impulse response is often very long (or infinitely long), high-order distributions of the input sequence would be required to determine the output CDF. In other words, if $Y[n]$ depends on the most recent k input values $X[n], \ldots, X[n-k+1]$, then the kth-order pdf of X has to be known in order to compute even the first-order pdf of Y. The situation with moment functions is different. The moments of the output random sequence can be calculated from equal- or lower-order moments of the input, when the system is linear. Partly for this reason, it is of considerable interest to determine the output moment functions in terms of the input moment functions. In the practical and important case of the Gaussian random sequence, we have seen that the entire probabilistic description depends only on the mean and covariance functions. In fact because the linear system is in effect performing a linear transformation on the infinite-dimensional vector that constitutes the input sequence, we can see that the output sequence will also obey the Gaussian law in its nth-order distributions if the input sequence is Gaussian. Thus, the determination of the first- and second-order moment functions of the output is particularly important when the input sequence is Gaussian.

Theorem 8.3-1 For a linear system L and a random sequence $X[n]$, the mean of the output random sequence $Y[n]$ is

$$E\{Y[n]\} = L\{E\{X[n]\}\} \tag{8.3-1}$$

as long as both sides are well defined.

Proof (formal). Since L is a linear operator, we can write

$$y[n] = \sum_{k=-\infty}^{+\infty} h[n, k]x[k]$$

for each sample sequence input–output pair, or

$$Y[n,\zeta] = \sum_{k=-\infty}^{+\infty} h[n,k]X[k,\zeta],$$

where we explicitly indicate the outcome ζ. If we operate on both sides with the expectation operator E, we get

$$E\{Y[n]\} = E\left\{\sum_{k=-\infty}^{+\infty} h[n,k]X[k]\right\}.$$

Now, assuming it is valid to bring the operator E inside the infinite sum, we get

$$\begin{aligned} E\{Y[n]\} &= \sum_{k=-\infty}^{+\infty} h[n,k]E\{X[k]\} \\ &= L\left\{E\{X[n]\}\right\}, \end{aligned}$$

which can be written as

$$\mu_Y[n] = \sum_{k=-\infty}^{+\infty} h[n,k]\mu_X[k],$$

that is, the mean function of the output is the response of the linear system to the mean function of the input. ■

Some comments are necessary with regard to this interchange of the expectation and linear operator. It cannot always be done! For example, if the input has a nonzero mean function and the linear system is a *running sum*, that is,

$$y[n] = \sum_{k=0}^{+\infty} x[n-k],$$

the running sum of the mean may not converge. Then such an interchange is not valid. We will come back to this point when we study stochastic convergence in Section 8.7. We will see then that a sufficient condition for an LSI system to satisfy Equation 8.3-1 is that its impulse response $h[n]$ be absolutely summable.

There are special cases of Equation 8.3-1 depending on whether the input sequence is WSS and whether the system is LSI. If the system is LSI and the input is at least WSS, then the mean of the output is given as

$$E\{Y[n]\} = \sum_{k=-\infty}^{+\infty} h[n-k]\mu_X.$$

Now because μ_X is a constant, we can take it out of the sum and obtain

$$E\{Y[n]\} = \left[\sum_{k=-\infty}^{+\infty} h[k]\right]\mu_X \tag{8.3-2}$$

$$= \mathsf{H}(z)|_{z=1}\,\mu_X, \tag{8.3-3}$$

at least whenever $\sum_{k=-\infty}^{+\infty} |h[k]|$ exists, that is, for any BIBO stable system (cf. Section 8.2).

Thus, we observe that in this case the mean of the output random sequence is a constant equal to the product of the *dc gain* or *constant gain* of the LSI system times the mean of the input sequence.

Example 8.3-1

(*lowpass filter*) Let the system be a *lowpass* filter with system function

$$\mathsf{H}(z) = 1/(1 + az^{-1}),$$

where we require $|a| < 1$ for stability of this assumed causal filter (i.e., the region of convergence is $|z| > |a|$, which includes the unit circle). Then if a WSS sequence is the input to this filter, the mean of the output will be

$$\begin{aligned} E\{Y[n]\} &= \mathsf{H}(z)|_{z=1} E\{X[n]\} \\ &= (1+a)^{-1}\mu_X. \end{aligned}$$

We now turn to the problem of calculating the output covariance and correlation of the general linear system whose operator is L:

$$Y[n] = L\{X[n]\}.$$

We will find it convenient to introduce a cross-correlation function between the input and output,

$$R_{XY}[m,n] \triangleq E\{X[m]Y^*[n]\} \tag{8.3-4}$$

$$= E\{X[m]\,(L\{X[n]\})^*\}. \tag{8.3-5}$$

Now, in order to factor out the operator, we introduce the operator L_n^*, with impulse response $h^*[n,k]$, which operates on time index k, but treats time index n as a constant. We can then write $R_{XY}[m,n] = E\{X[m]L_n^*[X^*[n]]\} = L_n^* E\{X[m]X^*[n]\}$. Similarly we denote with L_m the linear operator with time index m, that treats n as a constant. The operator L_n^* is related to the *adjoint operator* studied in linear algebra.

Theorem 8.3-2 Let $X[n]$ and $Y[n]$ be two random sequences that are the input and output, respectively, of the linear operator L_n. Let the input correlation function be $R_{XX}[m,n]$. Then the cross- and output-correlation functions are, respectively, given by

$$R_{XY}[m,n] = L_n^*\{R_{XX}[m,n]\}$$

and

$$R_{YY}[m,n] = L_m\{R_{XY}[m,n]\}.$$

Proof Write

$$
\begin{aligned}
X[m]Y^*[n] &= X[m]L_n^*\{X^*[n]\} \\
&= L_n^*\{X[m]X^*[n]\}.
\end{aligned}
$$

Then

$$
\begin{aligned}
R_{XY}[m,n] = E\{X[m]Y^*[n]\} &= E\{L_n^*\{X[m]X^*[n]\}\} \\
&= L_n^*\{E\{X[m]X^*[n]\}\} \\
&= L_n^*\{R_{XX}[m,n]\},
\end{aligned}
$$

thus establishing the first part of the theorem. To show the second part, we proceed analogously by multiplying $Y[m]$ by $Y^*[n]$ to get

$$
\begin{aligned}
E\{Y[m]Y^*[n]\} &= E\{L_m\{X[m]Y^*[n]\}\} \\
&= L_m\{E\{X[m]Y^*[n]\}\} \\
&= L_m\{R_{XY}[m,n]\},
\end{aligned}
$$

as was to be shown. ■

If we combine both parts of Theorem 8.3-2 we get an operator expression for the output correlation in terms of the input correlation function:

$$
R_{YY}[m,n] = L_m\{L_n^*\{R_{XX}[m,n]\}\}, \tag{8.3-6}
$$

which can be put into the form of a superposition summation for a system with time-variant impulse response $h[n,k]$ as

$$
R_{YY}[m,n] = \sum_{k=-\infty}^{+\infty} h[m,k]\left(\sum_{l=-\infty}^{+\infty} h^*[n,l]R_{XX}[k,l]\right). \tag{8.3-7}
$$

Here the superposition summation representation for $R_{XY}[m,n]$ is

$$
\begin{aligned}
R_{XY}[m,n] &= L_n^*\{R_{XX}[m,n]\} \\
&= \sum_{l=-\infty}^{+\infty} h^*[n,l]R_{XX}[m,l],
\end{aligned}
$$

and that for $R_{YX}[m,n]$ is

$$
R_{YX}[m,n] = \sum_{k=-\infty}^{+\infty} h[m,k]R_{XX}[k,n].
$$

To find the corresponding results for covariance functions, we note that the centered output sequence is the output due to the centered input sequence, due to the linearity of the system

and Equation 8.3-1. Then applying Theorem 8.3-2 to these zero-mean sequences, we have immediately that, for covariance functions,

$$K_{XY}[m,n] = L_n^*\{K_{XX}[m,n]\} \tag{8.3-8}$$

$$K_{YY}[m,n] = L_m\{K_{XY}[m,n]\} \tag{8.3-9}$$

and

$$K_{YY}[m,n] = L_m\{L_n^*\{K_{XX}[m,n]\}\}, \tag{8.3-10}$$

which becomes the following superposition summation

$$K_{YY}[m,n] = \sum_{k=-\infty}^{+\infty} h[m,k]\left(\sum_{l=-\infty}^{+\infty} h^*[n,l]K_{XX}[k,l]\right). \tag{8.3-11}$$

Example 8.3-2

(*edge detector*) Let $Y[n] \triangleq X[n] - X[n-1] = L\{X[n]\}$, an operator that represents a first-order (backward) difference. See Figure 10.3-1. This linear operator could be applied to locate an impulse noise spike in some random data. The output mean is $E[Y[n]] = L\{E[X[n]]\} = \mu_X[n] - \mu_X[n-1]$. The cross-correlation function is

$$\begin{aligned} R_{XY}[m,n] &= L_n\{R_{XX}[m,n]\} \\ &= R_{XX}[m,n] - R_{XX}[m,n-1]. \end{aligned}$$

The output autocorrelation function is

$$\begin{aligned} R_{YY}[m,n] &= L_m\{R_{XY}[m,n]\} \\ &= R_{XY}[m,n] - R_{XY}[m-1,n] \\ &= R_{XX}[m,n] - R_{XX}[m-1,n] - R_{XX}[m,n-1] + R_{XX}[m-1,n-1]. \end{aligned}$$

If the input random sequence were WSS with autocorrelation function,

$$R_{XX}[m,n] = a^{|m-n|}, \qquad 0 < a < 1,$$

then the above example would specialize to

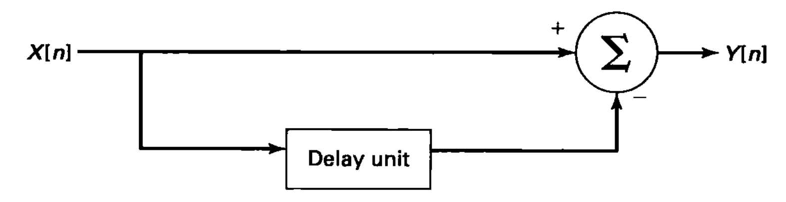


Figure 8.3-1 An edge detector that gives nearly zero output when $X[n] \approx X[n-1]$ and a large output when $|X[n] - X[n-1]|$ is large.

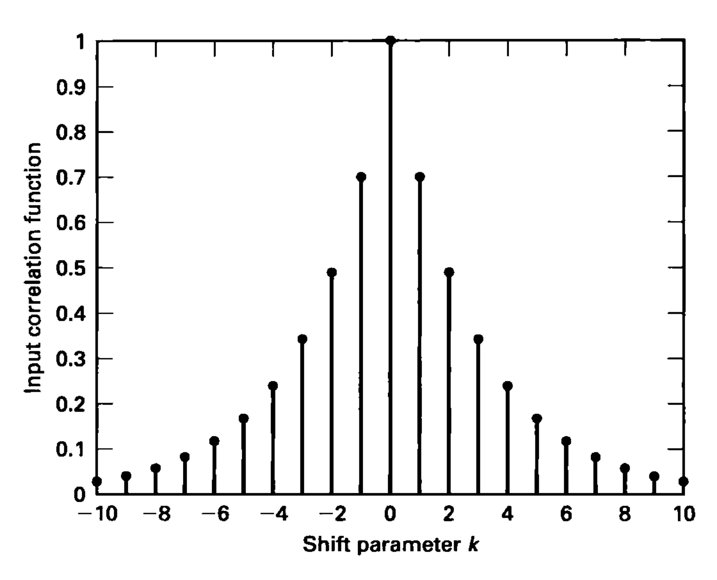


Figure 8.3-2 Input correlation function for edge detector with $a = 0.7$.

$$\mu_Y[n] = 0,$$

$$R_{XY}[m,n] = a^{|m-n|} - a^{|m-n+1|}$$

and

$$R_{YY}[m,n] = 2a^{|m-n|} - a^{|m-1-n|} - a^{|m-n+1|},$$

which depends on only $m - n$. Hence the output random sequence is WSS and we can write (with $k = m - n$)

$$R_{YY}[k] = 2a^{|k|} - a^{|k-1|} - a^{|k+1|}.$$

For the input autocorrelation with $a = 0.7$ as shown in Figure 8.3-2, the output autocorrelation function is shown in Figure 8.3-3. Note that the edge detector has a strong tendency to *decorrelate* the input sequence.

Example 8.3-3

(*covariance functions of a recursive system*) With $|\alpha| < 1$, let

$$Y[n] = \alpha Y[n-1] + (1-\alpha)W[n] \tag{8.3-12}$$

for $n \geq 0$ subject to $Y[-1] = 0$. Since the initial condition is zero, the system is equivalently LSI for $n \geq 0$, so we can represent L by convolution, where

$$h[n] = (1-\alpha)\alpha^n u[n].$$

Here $h[n]$ is the impulse response of the corresponding deterministic first-order difference equation, that is, $h[n]$ is the solution to

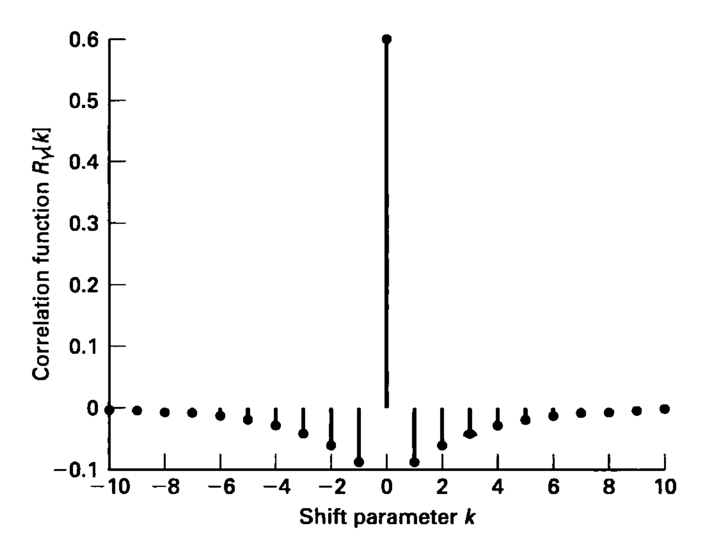


Figure 8.3-3 Correlation function $R_{YY}[k]$ for backward difference example (plot has $a = 0.7$).

$$h[n] = \alpha h[n-1] + (1-\alpha)\delta[n],$$

where $\delta[n]$ is the discrete-time impulse sequence. This solution can be obtained easily by recursion or by using the Z-transform.† Then specializing Equation 8.3-1, we obtain

$$\mu_Y[n] = \sum_{k=0}^{\infty}(1-\alpha)\alpha^k \mu_W[n-k], \text{ where } \mu_W[n] = 0 \text{ for } n < 0.$$

Applying Equations 8.3-8 and 8.3-9 to this case enables us to write, for α real,

$$K_{WY}[m,n] = \sum_{k=0}^{\infty}(1-\alpha)\alpha^k K_{WW}[m, n-k]$$

and

$$K_{YY}[m,n] = \sum_{l=0}^{\infty}(1-\alpha)\alpha^l K_{WY}[m-l, n],$$

which can be combined to yield

$$K_{YY}[m,n] = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}(1-\alpha)^2\alpha^k\alpha^l K_{WW}[m-l, n-k].$$

†Taking the Z-transform of both sides of the above equation, and noting that the Z-transform of the impulse sequence is 1, we obtain $\mathrm{H}(z) = (1-\alpha)/(1-\alpha z^{-1})$. Upon applying the inverse Z-transform, one gets the $h[n]$ given above. (For help with the inverse Z-transform, see Appendix A.)

Now if the input sequence $W[n]$ has covariance function

$$K_{WW}[m,n] = \sigma_W^2 \delta[m-n] \text{ for } m,n \geq 0,$$

then the output covariance is calculated as

$$\begin{aligned} K_{YY}[m,n] &= \sum_{k=0}^{n}(1-\alpha)^2\alpha^k\alpha^{(m-n)+k}\sigma_W^2 \quad \text{for} \quad m \geq n \geq 0, \\ &= \alpha^{(m-n)}(1-\alpha)^2\sum_{k=0}^{n}\alpha^{2k}\sigma_W^2 \\ &= \alpha^{(m-n)}\left[(1-\alpha)^2/(1-\alpha^2)\right]\sigma_W^2(1-\alpha^{2n+2}) \quad \text{for} \quad m \geq n \geq 0 \\ &= [(1-\alpha)/(1+\alpha)]\alpha^{|m-n|}\sigma_W^2(1-\alpha^{2\min(m,n)+2}) \quad \text{for all } m,n \geq 0, \end{aligned}$$

where the last step follows from the required symmetry in (m,n). Note that the term $\alpha^{2\min(m,n)+2}$ is a transient that dies away as $m,n \to \infty$, since $|\alpha| < 1$, so that asymptotically we have the *steady-state* answer

$$K_{YY}[m,n] = \left(\frac{1-\alpha}{1+\alpha}\right)\sigma_W^2\ \alpha^{|m-n|}, \quad m,n \to \infty,$$

a shift-invariant covariance function. If the mean function $\mu_Y[n]$ is found to be asymptotic to a constant, then the random sequence $Y[n]$ is said to be *asymptotically WSS*. We discuss WSS random sequences in greater detail in the next section.

As an alternative to this method of solution, one can take the expectation of Equation 8.3-12 to directly obtain a recursive equation for the output mean sequence which can be solved by the methods of Section 8.2:

$$\mu_Y[n] = \alpha\mu_Y[n-1] + (1-\alpha)\mu_W[n], \qquad n \geq 0,$$

with an appropriate initial condition. For example, if $\mu_Y[-1] = 0$ and $\mu_W[n] = \mu_W$, a given constant, then the solution is

$$\mu_Y[n] = (1-\alpha^{n+1})\mu_W u[n].$$

We can also use this method to calculate the cross-correlation function between input and output. First we conjugate Equation 8.3-12, then multiply by $W[m]$, and finally take the expectation to yield, for α real,

$$R_{WY}[m,n] = \alpha R_{WY}[m,n-1] + (1-\alpha)R_{WW}[m,n], \tag{8.3-13}$$

which can be solved directly for R_{WY} in terms of R_{WW}. The partial difference equation for the output correlation R_{YY} is obtained by re-expressing Equation 8.3-12 as a function of m, multiplying by $Y^*[n]$, and then taking the expectation to yield

$$R_{YY}[m,n] = \alpha R_{YY}[m-1,n] + (1-\alpha)R_{WY}[m,n]. \tag{8.3-14}$$

These two difference equations can be solved by the methods of Section 8.2 since they can each be seen to be one-dimensional difference equations with constant coefficients in one index, with the other index simply playing the role of an additional parameter. Thus, for example, one must solve Equation 8.3-13 as a function of n for each value of m in succession.

8.4 WSS RANDOM SEQUENCES

In this section we will assume that the random sequences of interest are all WSS, that is,

$$(1) \qquad E\{X[n]\} = \mu_X, \text{ a constant,}$$

$$(2) \quad R_{XX}[k+m,k] = E\{X[k+m]X^*[k]\} = R_{XX}[m],$$

and of *second order*, that is, $E\{|X[n]|^2\} < \infty$.

Some important properties of the autocorrelation function of stationary random sequences are presented below. They also hold for covariance functions, since they are just the autocorrelation function of the *centered* random sequence $X_c[n] \triangleq X[n] - \mu_X$.

1. For arbitrary m, $|R_{XX}[m]| \leq R_{XX}[0] \geq 0$, which follows directly from $E\{|X[m] - X[0]|^2\} \geq 0$ for $X[n]$ real valued, otherwise use Schwarz inequality (cf. Equation 4.3-15).
2. $|R_{XY}[m]| \leq \sqrt{R_{XX}[0]R_{YY}[0]}$, which is derived using the Schwarz inequality.
3. $R_{XX}[m] = R^*_{XX}[-m]$ since $R_{XX}[m] = E\{X[m+l]X^*[l]\} = E\{X[l]X^*[l-m]\} = E^*\{X[l-m]X^*[l]\} = R^*_{XX}[-m]$.
4. For all $N > 0$ and all complex $a_1, a_2, \ldots, a_N$, we must have

$$\sum_{n=1}^{N}\sum_{k=1}^{N} a_n a_k^* R_{XX}[n-k] \geq 0.$$

Property 4 is the *positive semidefinite* property of autocorrelation functions. It is a necessary and sufficient property for a function to be a valid autocorrelation function of a random sequence. In general it is very difficult to directly apply property 4 to test a function to see if it qualifies as a valid autocorrelation function. However, we soon will introduce an equivalent frequency domain function called *power spectral density*, which furnishes an easy test of validity.

Many of the input–output relations derived in the previous section take a surprisingly simple form in the case of WSS random sequences and LSI systems described via convolution. For example, starting with

$$Y[n] = \sum_{k=-\infty}^{+\infty} h[n-k]X[k],$$

we obtain

$$\begin{aligned}
R_{XY}[m,n] &= E\{X[m]Y^*[n]\} \\
&= \sum_{k=-\infty}^{+\infty} h^*[n-k]E\{X[m]X^*[k]\} \\
&= \sum_{k=-\infty}^{+\infty} h^*[n-k]R_{XX}[m-k] \\
&= \sum_{k=-\infty}^{+\infty} h^*[-l]R_{XX}[(m-n)-l], \quad \text{with } l \triangleq k-n,
\end{aligned}$$

if the input random sequence $X[n]$ is WSS. So, the output cross-correlation function $R_{XY}[m,n]$ is shift-invariant, and we can make use of the one-parameter cross-correlation function $R_{XY}[m] \triangleq R_{XY}[m,0]$ to write

$$\begin{aligned}
R_{XY}[m] &= \sum_{l=-\infty}^{+\infty} h^*[-l]R_{XX}[m-l] \\
&= h^*[-m] * R_{XX}[m],
\end{aligned}$$

in terms of the one-parameter autocorrelation function $R_{XX}[m]$. Likewise, recalling that

$$\begin{aligned}
R_{YY}[n+m,n] &\triangleq E\{Y[n+m]Y^*[n]\} \\
&= \sum_{k=-\infty}^{+\infty} h[k]E\{X[n+m-k]Y^*[n]\} \\
&= \sum_{k=-\infty}^{+\infty} h[k]R_{XY}[m-k] \\
&= h[m] * R_{XY}[m],
\end{aligned}$$

we see that the autocorrelation function of the output is shift-invariant, and so making use of the one-parameter autocorrelation function $R_{YY}[m] \triangleq R_{YY}[m,0]$, we have

$$R_{YY}[m] = h[m] * R_{XY}[m].$$

Combining both equations, we get

$$\begin{aligned}
R_{YY}[m] &= h[m] * h^*[-m] * R_{XX}[m] \\
&= (h[m] * h^*[-m]) * R_{XX}[m] \qquad (8.4\text{-}1) \\
&= g[m] * R_{XX}[m], \quad \text{with } g[m] \triangleq h[m] * h^*[-m]
\end{aligned}$$

where $g[m]$ is sometimes called the *autocorrelation impulse response* (AIR). Note that if the input random sequence is WSS and independent, then its autocorrelation function would be a positive constant times $\delta[m]$, so that taking this constant to be unity, we would have the output autocorrelation function equal to $g[m]$ itself. Therefore, $g[m]$ must possess all the properties of autocorrelation functions, that is , $g[l] = g^*[-l], g[0] \geq g[l]$ for all l, and positive semidefiniteness. The AIR g depends only on the impulse response h of the LSI system; however, in the absence of other information, we cannot uniquely determine h from g. In astronomy, crystallography, and other fields the problem of estimating h from the AIR is an important problem known by various names including *phase recovery* and *deconvolution*.

Example 8.4-1

(*impulse response*) We cannot in general calculate the impulse response from the AIR. To show this, first take the Fourier transform of $g[m]$ to obtain $G(\omega) = H(\omega)H^*(\omega) = |H(\omega)|^2$. Then note that $|H(\omega)| = \sqrt{G(\omega)}$. Thus the phase of $H(\omega)$ is lost in the AIR, but the magnitude of $H(\omega)$ is preserved. Often there is some information available that can narrow down or possibly pinpoint the phase, for example, the support of $h[n]$ in an image application, or causality for a time-based signal. For the interested reader, the literature contains many articles on this subject; see for example [8-6].

Example 8.4-2

(*correlation function analysis of the edge detector using impulse response*) In the edge detector of Example 8.3-2, the linear transformation was given as

$$Y[n] = L\{X[n]\} \triangleq X[n] - X[n-1],$$

an LSI operation with impulse response $h[n] = \delta[n] - \delta[n-1]$, and input autocorrelation function $R_{XX}[m] = a^{|m|}$, with $|a| < 1$. We can easily calculate the AIR as

$$\begin{aligned} g[m] &= h[m] * h[-m] \\ &= (\delta[m] - \delta[m-1]) * (\delta[-m] - \delta[-m-1]) \\ &= (\delta[m] - \delta[m-1]) * (\delta[m] - \delta[m+1]) \\ &= \delta[m] - \delta[m-1] - \delta[m+1] + \delta[m] \\ &= 2\delta[m] - \delta[m-1] - \delta[m+1]. \end{aligned}$$

We then calculate the output autocorrelation function in this WSS case as

$$\begin{aligned} R_{YY}[m] &= g[m] * R_{XX}[m] \\ &= (2\delta[m] - \delta[m-1] - \delta[m+1]) * a^{|m|} \\ &= 2a^{|m|} - a^{|m-1|} - a^{|m+1|}, \quad \text{for } -\infty < m < +\infty, \end{aligned}$$

which agrees with the answer in Example 8.3-2, where the result was plotted for $a = 0.7$.

Power Spectral Density

We define *power spectral density* (psd) as the FT (cf. Definition 8.2-3) of the one-parameter discrete-time autocorrelation function of a WSS random sequence $X[n]$:

$$S_{XX}(\omega) \triangleq \sum_{m=-\infty}^{+\infty} R_{XX}[m] \exp(-j\omega m), \quad \text{for } -\pi \leq \omega \leq +\pi. \tag{8.4-2}$$

Now by taking the FT of Equation 8.4-1, we get the following important psd input/output relation for an LSI system excited by a WSS random sequence:

$$S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega) = G(\omega) S_{XX}(\omega), \tag{8.4-3}$$

where the various frequency-domain quantities are discrete-time Fourier transforms. Equation 8.4-3 is a central result in the theory of WSS random sequences in that it enables the computation of the output psd directly from knowledge of the input psd and the transfer function magnitude. By using the IFT, we can calculate the autocorrelation function as

$$R_{XX}[m] = IFT\{S_{XX}(\omega)\} = \frac{1}{2\pi}\int_{-\pi}^{+\pi} S_{XX}(\omega) e^{j\omega m} d\omega,$$

so that knowledge of the psd implies knowledge of the autocorrelation function.

As to the name power spectral density, note that $R_{XX}[0] = E\{|X[n]|^2\}$ is the ensemble average power in $X[n]$ and so by the above relation, we see that

$$E\{|X[n]|^2\} = R_{XX}[0] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} S_{XX}(\omega) d\omega,$$

so that the integral average of the psd over its frequency range $[-\pi, +\pi]$ is indeed average power. To pursue this further, we consider a WSS random sequence $X[n]$ input to an LSI system consisting of a narrow band filter $H(\omega)$, with very small bandwidth 2ϵ, centered at frequency ω_o, where $|\omega_o| < \pi$, and with unity passband gain. Writing $S_{XX}(\omega)$ for the input psd, we have for the output ensemble average power, approximately

$$R_{YY}[0] = \frac{1}{2\pi}\int_{\omega_o-\epsilon}^{\omega_o+\epsilon} S_{XX}(\omega)\, d\omega \simeq S_{XX}(\omega_0)\frac{\epsilon}{\pi},$$

thus showing that $S_{XX}(\omega)$ can be interpreted as a density function in frequency for ensemble average power.

Some important properties of the psd are given below:

1. The function $S_{XX}(\omega)$ is real valued since $R_{XX}[m]$ is conjugate-symmetric.
2. If $X[n]$ is a real-valued random sequence, then $S_{XX}(\omega)$ is an even function of ω.
3. The function $S_{XX}(\omega) \geq 0$ for every ω, whether $X[n]$ is real- or complex-valued.
4. If $R_{XX}[m] = 0$ for all $|n| > N$ for some finite integer $N > 0$ (i.e., it has *finite support*), then $S_{XX}(\omega)$ is an *analytic function* in ω. This means that $S_{XX}(\omega)$ can be represented in a Taylor series given its value and that of all its derivatives at a single point ω_o.

Since $S_{XX}(\omega)$ is the Fourier transform of a (autocorrelation) sequence, it is periodic with period 2π. This is why the inverse Fourier transform, which recovers the autocorrelation function, only integrates over $[-\pi, +\pi]$, the *primary period.* We also define the Fourier transform of the cross-correlation function of two jointly WSS random sequences:

$$S_{XY}(\omega) \triangleq \sum_{m=-\infty}^{+\infty} R_{XY}[m]\exp(-j\omega m), \quad \text{for } -\pi \leq \omega \leq +\pi,$$

called the *cross-power spectral density* between random sequences X and Y. In general, this cross-power spectral density can be complex, negative, and lacking in symmetry. Its main use is as an intermediate step in calculation of psd's.

Interpretation of the psd

From its name, we expect that the psd should be related to some kind of average of the magnitude-square of the Fourier transform of the random signal. Now since a WSS random signal $X[n]$ has constant average power $R_{XX}[0]$ for all time, we cannot define its FT; however, we can define the transform quantity

$$X_N(\omega) \triangleq FT\{w_N[n]X[n]\}$$

with aid of the rectangular window function

$$w_N[n] \triangleq \begin{cases} 1, & |n| \leq N, \\ 0, & \text{else.} \end{cases}$$

Then, taking the expectation of the magnitude square $|X_N(\omega)|^2$, and dividing by $2N+1$, we get

$$\begin{aligned}
\frac{1}{2N+1}E\{|X_N(\omega)|^2\} &= \frac{1}{2N+1}E\left\{\sum_{k=-N}^{+N}\sum_{l=-N}^{+N} X[k]X^*[l]\exp(-j\omega k)\exp(+j\omega l)\right\} \\
&= \frac{1}{2N+1}\sum_{k=-N}^{+N}\sum_{l=-N}^{+N} E\{X[k]X^*[l]\}\exp(-j\omega k)\exp(+j\omega l) \\
&= \frac{1}{2N+1}\sum_{k=-N}^{+N}\sum_{l=-N}^{+N} R_{XX}[k-l]\exp[-j\omega(k-l)] \\
&= \sum_{m=-2N}^{+2N} R_{XX}[m]\left(1-\frac{|m|}{2N+1}\right)\exp(-j\omega m),
\end{aligned}$$

where the last line comes from the fact that $R_{XX}[k-l]$ is constant along diagonals $k-l=m$ of the $(2N+1)\times(2N+1)$ point square in the (k,l) plane.

Now as $N \to \infty$, the triangular function $(1-\frac{|m|}{2N+1})$ has less and less effect if $|R_{XX}[m]| \to 0$ as $|m| \to \infty$, as it must for the Fourier transform, that is, $S_{XX}(\omega)$ to exist. In fact,

if we assume that $|R_{XX}[m]|$ decays fast enough to satisfy $\sum_{m=-\infty}^{+\infty} |m||R_{XX}[m]| < \infty$, then we have

$$S_{XX}(\omega) = \lim_{N\to\infty} \frac{1}{2N+1} E\{|X_N(\omega)|^2\}. \tag{8.4-4}$$

In words we have that the ensemble average of the power at frequency ω in the windowed random sequence is given by the psd $S_{XX}(\omega)$. Note that we have said nothing about the *variance* of the random variable $\frac{1}{2N+1}|X_N(\omega)|^2$, but just that its *mean value* converges to the psd. In the study of spectral estimation (cf. Section 11.6), it is shown that the variance does not get small as N gets large, so that $\frac{1}{2N+1}|X_N(\omega)|^2$ *cannot* be considered a good estimate of the psd without first doing some averaging. In the language of statistics (Chapter 6) we say that $(2N+1)^{-1}|X_N(\omega)|^2$ is not a consistent estimator for $S_{XX}(\omega)$.

Example 8.4-3

Here is a MATLAB m-file to compute the psd's of the random sequences with memory in Example 8.1-16 for $p = 0.8$, 0.5, and 0.2.

```
function[psd1,psd2,psd3]=psdmarkov2(N,p1,p2,p3)}
mc1=0*ones(1,N);
mc2=0*ones(1,N);
mc3=0*ones(1,N);
for i=1:N
 mc1(i)=0.25*(((-1)*(2*p1-1))^(i-1));% The (-1)^(i-1) factor shifts the
spectrum to yield
 mc2(i)=0.25*(((-1)*(2*p2-1))^(i-1));%an even function of frequency.
Otherwise
 mc3(i)=0.25*(((-1)*(2*p3-1))^(i-1));%the highest frequency
components appear

end
x=linspace(-pi,pi,N);%at pi and the lowest at 2*pi.
psd1=abs(fft(mc1));
psd2=abs(fft(mc2));
psd3=abs(fft(mc3));
plot(x,psd1,x,psd2,x,psd3)
title('Power spectral density (psd) of random sequences with memory')
xlabel('radian frequency')
ylabel('psd value')
end
```

See the three plots in Figure 8.4-1.

Example 8.4-4

A stationary random sequence $X[n]$ has power spectral density $S_{XX}(\omega) = N_0 w(3\omega/4\pi)$, where the rectangular window function w is given as

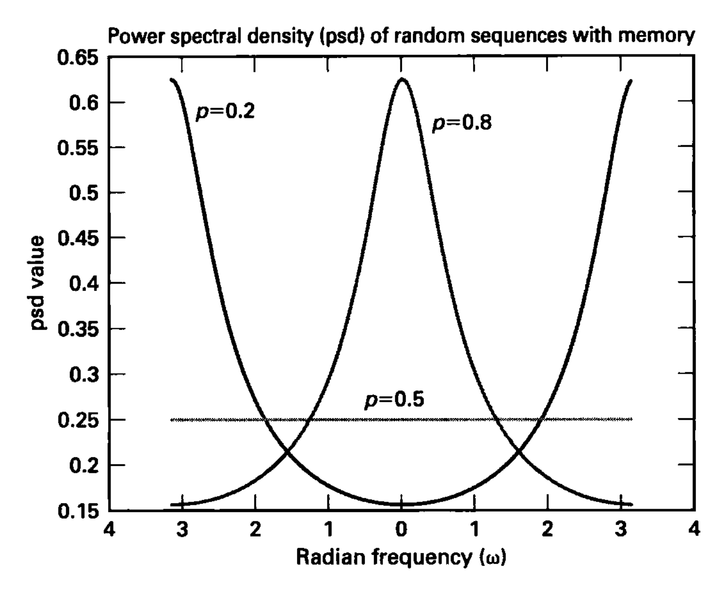


Figure 8.4-1 Power spectral densities of three stationary random sequences with memory.

$$w(x) \triangleq \begin{cases} 1, & |x| \leq 1/2, \\ 0, & \text{else.} \end{cases}$$

It is desired to produce an output random sequence $Y[n]$ with the psd $S_{YY}(\omega) = N_0 w(\omega/\pi)$. An LSI system (not necessarily causal) with impulse response $h[n]$ is proposed. Which of the following impulse responses should be used? (Note that $\text{sinc}(x) \triangleq \sin(\pi x)/\pi x$.)

(a) $2\text{sinc}(n/2)$,

(b) $\frac{1}{2}\text{sinc}((n-10)/2)$,

(c) $1.5e^{-|n|}u[n]$,

(d) $u[n+2] - u[n-2]$,

(e) $(1-|n|)w(n/2)$.

Solution Clearly what is needed is an $H(\omega)$ with transfer-function magnitude $|H(\omega)| = w(\omega/\pi)$. Choices (c) through (e) are ruled out immediately because their Fourier transforms do not have constant magnitude inside any frequency band. Since the IFT of $w(\omega/\pi)$ is $\frac{1}{2}\,\text{sinc}(n/2)$, we choose (b) since its 10-sample delay does not affect the magnitude $|H(\omega)|$.

A useful summary of input/output relations for random sequences is presented in Table 8.4-1.

Table 8.4-1 Input/Output Relations for WSS Sequences and Linear Systems

Random Sequence: $Y[n] = h[n] * X[n]$	Output Mean: $\mu_Y = H(0)\mu_X$
Crosscorrelations: $R_{XY}[m] = R_{XX}[m] * h^*[-m]$ $R_{YX}[m] = h[m] * R_{XX}[m]$ $R_{YY}[m] = R_{YX}[m] * h^*[-m]$	Cross–Power Spectral Densities: $S_{XY}(\omega) = S_{XX}(\omega)H^*(\omega)$ $S_{YX}(\omega) = H(\omega)S_{XX}(\omega)$ $S_{YY}(\omega) = S_{YX}(\omega)H^*(\omega)$
Autocorrelation: $R_{YY}[m] = h[m] * h^*[-m] * R_{XX}[m]$ $= g[m] * R_{XX}[m]$	Power Spectral Density: $S_{YY}(\omega) = \|H(\omega)\|^2 S_{XX}(\omega)$ $= G(\omega)\, S_{XX}(\omega)$

Output Power and Variance:

$$E\{|Y[n]|^2\} = R_{YY}[0] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} |H(\omega)|^2 S_{XX}(\omega)d\omega$$

$$\sigma_Y^2 = R_{YY}[0] - |\mu_Y|^2$$

Synthesis of Random Sequences and Discrete-Time Simulation

Here we consider the problem of finding the appropriate transfer function $H(\omega)$ to generate a random sequence with a specified psd or correlation function. Consider Equation 8.2-2, repeated here for convenience:

$$y[n] = \sum_{k=1}^{M} a_k y[n-k] + \sum_{k=0}^{N} b_k x[n-k], \tag{8.4-5}$$

where the coefficients are real valued. The transfer function $H(\omega)$ is given by

$$H(\omega) = \frac{Y(\omega)}{X(\omega)} = \frac{B(\omega)}{A(\omega)},$$

where $B(\omega) \triangleq \sum_{k=0}^{N} b_k e^{-j\omega k}$ and $A(\omega) \triangleq 1 - \sum_{k=1}^{M} a_k e^{-j\omega k}$. When driven by a white-noise sequence, $W[n]$, with power $E\{|W[n]|^2\} = \sigma_W^2$, the output psd, $S_{YY}(\omega)$, is given by

$$S_{YY}(\omega) = \frac{|B(\omega)|^2}{|A(\omega)|^2}\sigma_W^2 = \frac{B(\omega)B^*(\omega)}{A(\omega)A^*(\omega)}\sigma_W^2. \tag{8.4-6}$$

Now, recalling that $\mathsf{B}(z) \triangleq B(\omega)$ at $z = e^{j\omega}$ and similarly for $\mathsf{A}(z)$, $\mathsf{H}(z)$, that $e^{-j\omega} = z^{-1}$, and that $\mathsf{B}^*(e^{j\omega}) = \mathsf{B}(e^{-j\omega})$, we obtain an LCCDE with real coefficients†

†Only when, as here, the impulse response coefficients are real valued. This is true here since the numerator and denominator coefficients are real numbers.

$$S_{YY}(z) = \frac{B(z)B(z^{-1})}{A(z)A(z^{-1})}\sigma_W^2 = H(z)H(z^{-1})\sigma_W^2, \tag{8.4-7}$$

where up to this point we have confined z to the *unit circle*. For the purpose of further analysis, it is of interest to extend Equation 8.4-7 to the whole *z-plane*.

This last step is called *analytic continuation* and simply amounts to finding a rational function of z which agrees with the given psd information on the unit circle $z = e^{j\omega}$.

Given any rational $S_{XX}(z)$, that is, one with a finite number of poles and zeros in the finite z-plane, one can find such a *spectral factorization* as Equation 8.4-7 by defining $H(z)$ to have all the poles and zeros that are inside the unit circle, $\{|z| < 1\}$, and then $H(z^{-1})$ will necessarily have all the poles and zeros outside the unit circle, $\{|z| > 1\}$.

Example 8.4-5

Consider the psd

$$S_{XX}(\omega) = \frac{\sigma_W^2}{1 - 2\rho\cos\omega + \rho^2} \qquad \text{with } |\rho| < 1.$$

We want to first extend $S_{XX}(\omega)$ to all of the z-plane. Now $\cos\omega = \frac{1}{2}(e^{+j\omega} + e^{-j\omega})$, which can be extended as $\frac{1}{2}(z + z^{-1})$ and satisfies the symmetry condition $S_{XX}(z) = S_{XX}(z^{-1})$ of a real-valued random sequence. Then

$$\begin{aligned} S_{XX}(z) &= \frac{\sigma_W^2}{1 - \rho(z + z^{-1}) + \rho^2} \qquad \text{for } |\rho| < |z| < \frac{1}{|\rho|}, \\ &= \frac{\sigma_W^2}{(1 - \rho z)(1 - \rho z^{-1})} \\ &= \sigma_W^2 H(z)H(z^{-1}) \end{aligned}$$

with $H(z) = \dfrac{1}{1 - \rho z^{-1}}$ for region of convergence $|\rho| < |z|$.

Since $|\rho| < 1$, the region of convergence (ROC) includes the unit circle and so H is both stable and causal. Indeed the system with $h[n] = \rho^n u[n]$ will yield $S_{XX}(\omega)$ from an independent sequence.

If a zero occurs on the unit circle, then it must be of even order, since otherwise one can easily show that $S_{XX}(e^{j\omega})$ must go through zero and hence be negative in its vicinity. Thus, we can assign half the zeros to $H(z)$ and the other half to $H(z^{-1})$. Since $H(z)$ contains only poles inside the unit circle, it will be BIBO stable [8-5]. Except in the case of a zero on the unit circle, its inverse will also be stable. The other factor $H(z^{-1})$ has all its poles outside the unit circle, so it is stable in the anticausal sense. Denoting the largest pole magnitude inside the unit circle by $p_{\max}$, we thus have that $S_{XX}(z)$ is analytic, that is, free of singularities in the annular region of convergence $\{p_{\max} < |z| < 1/p_{\max}\}$.

Following the above procedures, we can obtain the system function $H(z)$ that, when driven by a white noise $W[n]$, will generate a random sequence $X[n]$ with special psd

$S_{XX}(\omega)$. This can be the basis for a discrete-time simulation on a computer. The white random sequence $W[n]$ is easily obtained by using the computer's random number generator. Then one specifies appropriate initial conditions and proceeds to recursively calculate $X[n]$ using the LCCDE of the system function $\mathsf{H}(z)$.

To achieve a Gaussian distribution for X, one could transform the output of the random number generator to achieve a Gaussian distribution for W, which would carry across to X. An approximate method that is often used is to average six to ten calls to the random number generator to obtain an approximate Gaussian distribution for W via the Central Limit theorem. When simulating a non-Gaussian random variable, the distribution for X and W is not the same. Thus the preceding method will not work. One possibility is to use the LCCDE to generate samples of $W[n]$ from some real data and then use the resulting distribution for $W[n]$ in the simulation.

Example 8.4-6

(*matching given correlation values*) In order to simulate a zero-mean random sequence with average power $R_{XX}[0] = \sigma^2$ and nearest neighbor correlation $R_{XX}[1] = \rho\sigma^2$, we want to find the parameters of a first-order stochastic difference equation to achieve these values. Thus consider

$$X[n] = aX[n-1] + bW[n], \tag{8.4-8}$$

where $W[n]$ is a zero-mean white-noise source with unit power. Computing the impulse response, we get

$$h[n] = ba^n u[n]$$

and the corresponding system function

$$\mathsf{H}(z) = \frac{b}{1 - az^{-1}}.$$

Since the mean is zero, we calculate the covariance of the output $X[n]$ of Equation 8.4-8:

$$\begin{aligned}
K_{XX}[m] &= h[m] * h[-m] * K_{WW}[m] \\
&= h[m] * h[-m] \\
&= b^2 \left(a^m u[m]\right) * \left(a^{-m} u[-m]\right) \\
&= b^2 \sum_{k=-\infty}^{+\infty} a^k u[k] a^{m+k} u[m+k] \\
&= b^2 a^m \sum_{k=\max(0,-m)}^{+\infty} a^{2k} \\
&= \frac{b^2}{1-a^2} a^{|m|}, \qquad -\infty < m < +\infty.
\end{aligned}$$

From the specifications at $m = 0$ and $m = 1$, we need

$$K_{XX}[0] = \sigma^2 = b^2/(1 - a^2),$$
$$K_{XX}[1] = \rho\sigma^2 = ab^2/(1 - a^2).$$

Thus,

$$a = \rho \text{ and } b^2 = \sigma^2(1 - \rho^2).$$

To compute the resulting psd, we use Equation 8.4-4 to get

$$\begin{aligned} S_{XX}(\omega) &= \frac{b^2}{|1 - ae^{-j\omega}|^2} \\ &= \frac{\sigma^2(1 - \rho^2)}{1 - 2\rho\cos\omega + \rho^2}. \end{aligned}$$

Example 8.4-7

(*decimation and interpolation*) Let $X[n]$ be a WSS random sequence. We consider what happens to its stationarity and psd when we subject it to decimation or interpolation as occur in many signal processing systems.

Decimation

Set $Y[n] \triangleq X[2n]$, called *decimation* by the factor 2, thus throwing away every odd indexed sample of $X[n]$ (Figure 8.4-2). We easily calculate the mean function as $\mu_Y[n] \triangleq E\{Y[n]\} = E\{X[2n]\} = \mu_X[2n] = \mu_X$, a constant. For the correlation,

$$\begin{aligned} R_{YY}[n + m, n] &= E\{X[2n + 2m]X^*[2n]\} \\ &= R_{XX}[2n + 2m, 2n] \\ &= R_{XX}[2m], \end{aligned}$$

thus showing that the WSS property of the original random sequence $X[n]$ is preserved in the decimated random sequence. The psd of $Y[n]$ can be computed as

$$\begin{aligned} S_{YY}(\omega) &\triangleq \sum_{m=-\infty}^{+\infty} R_{YY}[m]\exp[-j\omega m] \\ &= \sum_{m=-\infty}^{+\infty} R_{XX}[2m]\exp[-j\omega m] \\ &= \sum_{m \text{ even}} R_{XX}[m]\exp\left[-j\frac{\omega}{2}m\right] = \sum_{m \text{ even}} R_{XX}[m]\exp\left[-j\frac{\omega}{2}m\right](-1)^m. \end{aligned}$$

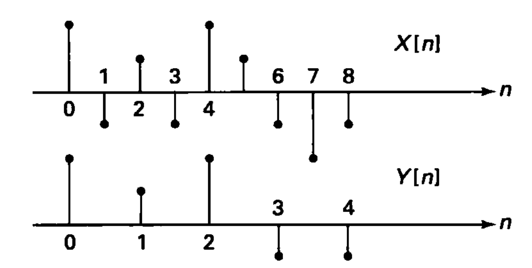


Figure 8.4-2 In decimation every other value of $X[n]$ is discarded.

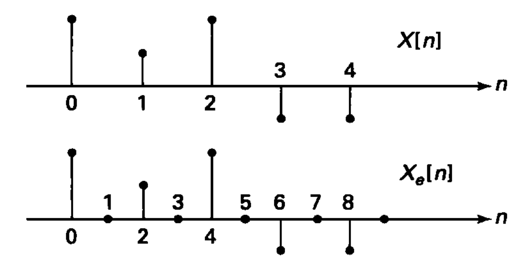


Figure 8.4-3 In interpolation, the expansion step inserts zeros between adjacent values of the $X[n]$ sequence, to get the expanded sequence $X_e[n]$.

Now, define $A_e \triangleq S_{YY}(\omega)$, and $A_o \triangleq \sum_{m \text{ odd}} R_{XX}[m] \exp[-j\frac{\omega}{2}m]$. Then, clearly $A_e + A_o = S_{XX}(\frac{\omega}{2})$ and $A_e - A_o = S_{XX}(\frac{\omega-2\pi}{2})$, so that

$$S_{YY}(\omega) = \frac{1}{2}\left[S_{XX}\left(\frac{\omega}{2}\right) + S_{XX}\left(\frac{\omega - 2\pi}{2}\right)\right],$$

which displays an *aliasing* [8-5] of higher-frequency terms.

Interpolation

For interpolation by the factor 2, we do the opposite of decimation. First we perform an *expansion* by setting

$$X_e[n] \triangleq \begin{cases} X[\frac{n}{2}], & n = \text{even} \\ 0, & n = \text{odd}. \end{cases}$$

The resulting expanded random sequence is clearly nonstationary, because of the zero insertions. See Figure 8.4-3. Formally the psd of $X_e[n]$ doesn't exist since the psd is defined only for WSS sequences (Figure 8.4-4). We encounter such problems with a broad class of random sequences and processes[†] classified as being *cyclostationary* (cf. Section 9.6) to which $X_e[n]$ belongs. It is easy to convert such sequences to WSS by randomizing their start times and then averaging over the start time (Example 9.6-1). However, here we instead compute the power spectral density using Equation 8.4-4, which is permissible for cyclostationary waveforms. Thus we write

[†] *Random processes* are continuous-time random waveforms to be discussed in Chapter 9.

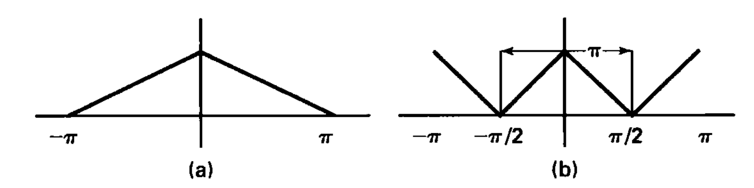


Figure 8.4-4 (a) The original psd of $X[n]$; (b) the psd of $X_e[n]$ (not drawn to scale). Note the "leakage" of power density from the secondary periods into the primary period. An ideal lowpass filter with support $[-\pi/2, \pi/2]$ will eliminate the contribution from the secondary periods.

$$E\{|Y_N(\omega)|^2\} = E\left\{\left|\sum_{n=-N}^{N} X_e[n]e^{-j\omega n}\right|^2\right\}$$

and take the limit of $E\{|X_e^{(N)}(\omega)|^2\}/(2N+1)$ as $N \to \infty$. This quantity can be interpreted as the psd, $S_{X_eX_e}(\omega)$, of the random sequence $X_e[n]$. If the algebra is carried out and we assume that $R_{XX}[m]$ is absolutely summable, we find that $S_{X_eX_e}(\omega) = \frac{1}{2}S_{XX}(2\omega)$. For further discussion of the expansion step, see Problems 8.58 and 8.59.

Next we put $X_e[n]$, sometimes called an *upsampled version* of $X[n]$, through an ideal lowpass filter with bandwidth $[-\frac{\pi}{2}, +\frac{\pi}{2}]$ and gain of 2, to produce the "ideal" interpolated output $Y[n]$ as

$$Y[n] \triangleq h[n] * X_e[n].$$

The impulse response of such a filter is

$$h[n] = \frac{\sin(\pi n/2)}{(\pi n/2)}.$$

Thus,

$$\begin{aligned} Y[n] &= \sum_{k=-\infty}^{+\infty} X_e[k]\frac{\sin(n-k)\pi/2}{(n-k)\pi/2} \\ &= \sum_{k=-\infty}^{+\infty} X[k]\frac{\sin(n-2k)\pi/2}{(n-2k)\pi/2}. \end{aligned}$$

First we calculate the mean function of $Y[n]$,

$$\begin{aligned} \mu_Y[n] &\triangleq E\{Y[n]\} \\ &= E\left\{\sum_{k=-\infty}^{+\infty} X[k]\frac{\sin(n-2k)\pi/2}{(n-2k)\pi/2}\right\} \\ &= \sum_{k=-\infty}^{+\infty} \mu_X[k]\frac{\sin(n-2k)\pi/2}{(n-2k)\pi/2} \\ &= \mu_X \sum_{k=-\infty}^{+\infty} \frac{\sin(n-2k)\pi/2}{(n-2k)\pi/2}, \end{aligned}$$

the last step being allowed since μ_X is a constant. Now sampling theory can be used to show that the infinite sum is 1, so that $\mu_Y[n] = \mu_X$. To see this we write the sampling theorem representation for an arbitary bandlimited function $g(t)$ and sampling period $T = 2$ as in [8-5]:

$$g(t) = \sum_{k=-\infty}^{+\infty} g(2k)\frac{\sin(t-2k)\pi/2}{(t-2k)\pi/2}. \tag{8.4-9}$$

Then we simply choose $t = n$ for the bandlimited function $g(t) = 1$ with zero bandwidth to see that

$$1 = \sum_{k=-\infty}^{+\infty} \frac{\sin(n-2k)\pi/2}{(n-2k)\pi/2}.$$

For future reference we define $h(t) \triangleq \frac{\sin \pi t/2}{\pi t/2}$. To find the correlation function, we proceed to calculate

$$\begin{aligned}
R_{YY}[n+m,n] &= E\{Y[n+m]Y^*[n]\} \\
&= \sum_{k_1,k_2=-\infty}^{+\infty} E\{X[k_1]X^*[k_2]\}h[n+m-2k_1][h[n-2k_2] \\
&= \sum_{l_1=\text{even}} R_{XX}[l_1] \sum_{l_2=\text{even}} h[n+m-l_1-l_2]h[n+l_1-l_2] \\
&\quad + \sum_{l_1=\text{odd}} R_{XX}[l_1] \sum_{l_2=\text{odd}} h[n+m-l_1-l_2]h[n+l_1-l_2]
\end{aligned}$$

with $l_1 \triangleq k_1 - k_2$ and $l_2 \triangleq k_1 + k_2$ and $l_2 \pm l_1$ even. We can evaluate the sums

$$\sum_{l_2=\text{even or odd}} h[n+m-l_1-l_2]h[n+l_1-l_2]$$

by letting $g(t) = h(t)$ in Equation 8.4-9 and allowing t to take the value $t = m$. We find that each sum, both the even and odd, equals $h[m-2l_1]$. Thus,

$$R_{YY}[m+n,n] = R_{YY}[m] = \sum_{l_1} R_{XX}[l_1]h[m-2l_1].$$

We thus see that $Y[n]$ is WSS, that $R_{YY}[m]$ interpolates $R_{XX}[m]$, that is,

$$\begin{aligned}
R_{YY}[2m] &= \sum_{l_1} R_{XX}[l_1]h[2m-2l_1] \\
&= R_{XX}[2m],
\end{aligned}$$

and calculating the psd

$$\begin{aligned}
S_{YY}(\omega) &= \sum_m R_{YY}[m]e^{-j\omega m} \\
&= \sum_m \sum_{l_1} R_{XX}[l_1]h[m-2l_1]e^{-j\omega m} \\
&= \sum_{l_1} R_{XX}[l_1] \sum_m h[m-2l_1]e^{-j\omega m} \\
&= \sum_{l_1} R_{XX}[l_1]H(\omega)e^{-j2\omega l_1} \\
&= H(\omega)S_{XX}(2\omega) \\
&= \begin{cases} 2S_{XX}(2\omega), & |\omega| \leq \pi/2, \\ 0, & \frac{\pi}{2} < |\omega| \leq \pi \end{cases}.
\end{aligned}$$

8.5 MARKOV RANDOM SEQUENCES

We have already encountered some examples of Markov random sequences. Such sequences were loosely said to have a *memory* and to possess a *state*. Here we make these concepts more precise. We start with a definition.

Definition 8.5-1 (*Markov random sequence*)

(a) A *continuous-valued Markov random sequence* $X[n]$, defined for $n \geq 0$, satisfies the conditional pdf expression

$$f_X(x_{n+k}|x_n, x_{n-1}, \ldots, x_0) = f_X(x_{n+k}|x_n)$$

for all $x_0, \ldots, x_n, x_{n+k}$, for all $n > 0$, and for all integers $k \geq 1$.

(b) A *discrete-valued Markov random sequence* $X[n]$, defined for $n \geq 0$, satisfies the conditional PMF expression

$$P_X(x_{n+k}|x_n, \ldots, x_0) = P_X(x_{n+k}|x_n)$$

for all $x_0, \ldots, x_n, x_{n+k}$, for all $n > 0$, and for all $k \geq 1$. ■

It is sufficient for the above properties to hold for just $k = 1$, which is the so-called *one-step* case, as the general property can be built up from it. The *discrete-valued* Markov random sequence is also called a Markov chain and will be covered in the next section. Here we consider the continuous-valued case.

To check the meaning and usefulness of the Markov concept, consider the general Nth-order pdf $f_X(x_N, x_{N-1}, \ldots, x_0)$ of random sequence X, and repeatedly use conditioning to obtain the *chain rule* of probability

$$f_X(x_0, x_1, \ldots, x_N) = f_X(x_0)f_X(x_1|x_0)f_X(x_2|x_1, x_0)\ldots f_X(x_N|x_{N-1}, \ldots, x_0). \quad (8.5\text{-}1)$$

Now substitute the basic one-step ($k = 1$) version of the Markov definition to obtain

$$f_X(x_0, x_1, \ldots, x_N) = f_X(x_0) f_X(x_1|x_0) f_X(x_2|x_1) \ldots f_X(x_N|x_{N-1})$$
$$= f_X(x_0) \prod_{k=1}^{N} f_X(x_k|x_{k-1}).$$

Next we present two examples of continuous-valued Markov random sequences which are Gaussian distributed.

Example 8.5-1

(*Gauss Markov random sequence*) Let $X[n]$ be a random sequence defined for $n \geq 1$, with initial pdf

$$f_X(x;0) = N(0, \sigma_0^2)$$

for a given $\sigma_0 > 0$ and transition pdf

$$f_X(x_n|x_{n-1}; n, n-1) \sim N(\rho x_{n-1}, \sigma_W^2)$$

with $|\rho| < 1$ and $\sigma_W > 0$. We want to determine the unconditional density of $X[n]$ at an arbitrary time $n \geq 1$ and proceed as follows.

In general, one would have to advance recursively from the initial density by performing the integrals (cf. Equation 2.6-84)

$$f_X(x;n) = \int_{-\infty}^{+\infty} f_X(x|\xi; n, n-1) f_X(\xi; n-1) d\xi \tag{8.5-2}$$

for $n = 1, 2, 3,$ and so forth. However, in this example we know that the unconditional first-order density will be Gaussian because each of the pdf's in Equation 8.5-2 is Gaussian, and the Gaussian density "reproduces itself" in this context; that is, the product of two exponential functions is still exponential. Hence the pdf $f_X(x;n)$ is determined by its first two moments. We first calculate the mean function

$$\mu_X[n] = E\{X[n]\}$$
$$= E[E\{X[n]|X[n-1]\}]$$
$$= E[\rho X[n-1]]$$
$$= \rho\mu_X[n-1],$$

where the outer expectation is over the values of $X[n-1]$. We thus obtain the recursive equation

$$\mu_X[n] = \rho\mu_X[n-1], \qquad n \geq 1,$$

with prescribed initial condition $\mu_X[0] = 0$. Hence $\mu_X[n] = 0$ for all n.

We also need the variance function $\sigma_X^2[n]$, which in this case is just $E[X^2[n]]$ since the mean is zero. Calculating, we obtain

$$
\begin{aligned}
E\{X^2[n]\} &= E[E\{X^2[n]|X[n-1]\}] \\
&= E[\sigma_W^2 + \rho^2 X^2[n-1]] \\
&= \sigma_W^2 + \rho^2 E\{X^2[n-1]\}
\end{aligned}
$$

or

$$
\sigma_X^2[n] = \rho^2 \sigma_X^2[n-1] + \sigma_W^2, \qquad n \geq 1.
$$

This is a first-order difference equation, which can be solved for $\sigma_X^2[n]$ given the condition $\sigma_X^2[0] = \sigma_0^2$ supplied by the initial pdf. The solution then is

$$
\begin{aligned}
\sigma_X^2[n] &= [1 + \rho^2 + \rho^4 + \ldots + \rho^{2(n-1)}]\sigma_W^2 + \rho^{2n}\sigma_0^2 \\
&\to \frac{1}{1-\rho^2}\sigma_W^2 \quad \text{as } n \to \infty.
\end{aligned}
$$

Example 8.5-2

(*Markov difference equation*) Consider the difference equation

$$
X[n] = \rho X[n-1] + W[n],
$$

where $W[n]$ is an independent random sequence (cf. Definition 8.1-2). Let $n > 0$; then

$$
\begin{aligned}
f_X(x_n, x_{n-1}, \ldots, x_0) &= f_X(x_n|x_{n-1}) f_X(x_{n-1}|x_{n-2}) \ldots f_X(x_1|x_0) f_W(x_0) \\
&= \left(\prod_{k=1}^{n} f_W(x_k - \rho x_{k-1}) \right) f_W(x_0),
\end{aligned}
$$

where $x[n] = x_n$ and $w[n] = w_n$ are the sample function values taken on by the random sequences $X[n]$ and $W[n]$, respectively. Clearly $X[n]$ is a Markov random sequence. If $W[n]$ is an independent and Gaussian random sequence, then this is just the case of Example 8.5-1 above. Otherwise, the Markov sequence $X[n]$ will be non-Gaussian.

The Markov property can be generalized to cover higher-order dependence and higher-order difference equations, thus extending the *direct dependence* concept to more than one-sample distance.

Definition 8.5-2 (*Markov-p random sequence*) Let the positive integer p be called the *order* of the Markov-p random sequence. A continuous-valued Markov-p random sequence $X[n]$, defined for $n \geq 0$, satisfies the conditional pdf equations

$$
f_X(x_{n+k}|x_n, x_{n-1}, \ldots, x_0) = f_X(x_{n+k}|x_n, x_{n-1}, \ldots, x_{n-p+1})
$$

for all $k \geq 1$ and for all $n \geq p$. ■

Returning to look at Equation 8.5-1, we can see that as the Markov order p increases, the modeling error in approximating a general random sequence by a Markov random sequence should get better.

$$
\begin{aligned}
f_X(x_0, x_1, \ldots, x_N) \\
&= f_X(x_0) f_X(x_1|x_0) f_X(x_2|x_1, x_0) \ldots f_X(x_N|x_{N-1}, \ldots, x_0) \\
&\approx f_X(x_0) f_X(x_1|x_0) f_X(x_2|x_1, x_0) \ldots f_X(x_p|x_{p-1}, \ldots, x_0) \\
&\qquad \times \prod_{k=p+1}^{N} f_X(x_k|x_{k-1}, \ldots, x_{k-p+1}).
\end{aligned}
$$

This approximation would be expected to hold for the usual case where the strongest dependence is on the nearby values, say $X[n-1]$ and $X[n-2]$, with the conditional dependence on far away values being generally negligible. When the Markov-p model is used in signal processing, one of the most important issues is determining an appropriate model order p so that statistics like the joint pdf's (Equation 8.5-1) of the *original data* are adequately approximated by those of the Markov-p model. In Chapter 11 on applications in statistical signal processing, we will see that Markov-p random sequences are quite useful in modern spectral estimation. The celebrated Kalman filter for the recursive linear estimation of distorted signals in noise is based on the Markov models.

ARMA Models

A class of linear constant coefficient difference equation models are called ARMA for *autoregressive moving average.* Here the input is an independent random sequence $W[n]$ with mean $\mu_W = 0$ and variance $\sigma_W^2 = 1$. The LCCDE model then takes the form

$$X[n] = \sum_{k=1}^{M} a_k X[n-k] + \sum_{k=0}^{L} b_k W[n-k].$$

If the model is BIBO stable and $-\infty < n < +\infty$, then a WSS output sequence results with psd

$$S_{XX}(\omega) = \frac{\left| \sum_{k=0}^{L} b_k \exp(-j\omega k) \right|^2}{\left| 1 - \sum_{k=1}^{M} a_k \exp(-j\omega k) \right|^2}.$$

The ARMA sequence is not Markov, but when $L = 0$, the sequence is Markov-M, and the resulting model is called *autoregressive* (AR). On the other hand when $M = 0$, that is, there are no feedback coefficients c_k, the equation becomes just

$$X[n] = \sum_{k=0}^{L} d_k W[n-k],$$

and the model is called *moving average* (MA). The MA model is often used to estimate the time-average value over a data window, as shown in the next example.

Example 8.5-3

(*running time average*) Consider a sequence of independent random variables $W[n]$ on $n \geq 1$. Denote their running time average as

$$\widehat{\mu}_W[n] = \frac{1}{n}\sum_{k=1}^{n} W[k].$$

Since we can write $\widehat{\mu}_W[n]$ equivalently as satisfying the time-varying AR equation,

$$\widehat{\mu}_W[n] = \frac{n-1}{n}\widehat{\mu}_W[n-1] + \frac{1}{n}W[n],$$

it follows from the joint independence of the input $W[n]$ that $\widehat{\mu}_W[n]$ is a nonstationary Markov random sequence.[†]

Markov Chains

A Markov random sequence can take on either continuous or discrete values and then be represented either by probability density functions (pdf's) or probability mass functions (PMFs) accordingly. In the discrete-valued case, we call the random sequence a Markov chain. Applications occur in buffer occupancy, computer networks, and discrete-time approximate models for the continuous-time Markov chains (cf. Chapter 9).

Definition 8.5-3 (*Markov chain*) A discrete-time Markov chain is a random sequence $X[n]$ whose Nth-order conditional PMFs satisfy

$$P_X(x[n]|x[n-1], \ldots, x[n-N]) = P_X(x[n]|x[n-1]) \tag{8.5-3}$$

for all n, for all values of $x[k]$, and for all integers $N > 1$. ■

The value of $X[n]$ at time n is called "*the state.*" This is because this current value, that is, the value at time n, determines future conditional PMFs, independent of the past values taken on by $X[n]$.

A practical case of great importance is when the range of values taken on by $X[n]$ is finite, say M. The discrete range of $X[n]$, that is, the values that X takes on, is sometimes referred to as a set of *labels*. The usual choices for the label set are either the integers $\{1, M\}$, or $\{0, M-1\}$. Such a Markov chain is said to have a finite *state space*, or is simply a *finite-state Markov chain*. In this case, and when the random sequence is stationary, we can represent the statistical transition information in a matrix $\mathbf{P}$ with entries

$$p_{ij} = P_{X[n]|X[n-1]}(j|i), \tag{8.5-4}$$

for $1 \leq i, j \leq M$. The matrix $\mathbf{P}$ is referred to as the *state-transition matrix*. Its defining property is that it is a matrix with nonnegative entries, whose rows sum to 1. Usually, and

[†]Note that the variance of $\widehat{\mu}_W[n]$ decreases with n.

again without loss of generality, we can consider that the Markov chain starts at time index $n = 0$. Then we must specify the set of initial probabilities of the states at $n = 0$, that is, $P_X(i;0), 1 \leq i \leq M$, which can be stored in the initial probability vector $\mathbf{p}[0]$, a *row vector* with elements $(\mathbf{p}[0])_i = P_X(i;0), 1 \leq i \leq M$.

The following example re-introduces the useful concept of *state-transition diagram*, already seen in Example 8.1-15.

Example 8.5-4

(*two-state Markov chain*) Let $M = 2$; then we can summarize transition probability information about a two-state Markov chain in Figure 8.5-1. The only addition needed is the set of initial probabilities, $P_X(1;0)$ and $P_X(2;0)$.

Possible questions might be: Given that we are in state 1 at time 4, what is the probability we end up in state 2 at time 6? Or given a certain probability distribution over the two states at time 3, what is the probability distribution over the two states at time 7? Note that there are several ways or paths to go from one state at one time to another state several time units later. The answers to these questions thus will involve a summation over these mutually exclusive outcomes.

Here we have $M = 2$, and the two-element probability *row vector* $\mathbf{p}[n] = (p_0[n], p_1[n])$. Using the state-transition matrix, we then have

$$\begin{aligned}\mathbf{p}[1] &= \mathbf{p}[0]\mathbf{P} \\ \mathbf{p}[2] &= \mathbf{p}[1]\mathbf{P} \\ &= \mathbf{p}[0]\mathbf{P}^2\end{aligned}$$

or, in general,

$$\mathbf{p}[n] = \mathbf{p}[0]\mathbf{P}^n.$$

In a statistical steady state, if one exists, we would have

$$\mathbf{p}[\infty] = \mathbf{p}[\infty]\mathbf{P}, \text{ where } \mathbf{p}[\infty] = \lim_{n\to\infty} \mathbf{p}[n].$$

Writing $\mathbf{p} \triangleq \mathbf{p}[\infty]$, we have $\mathbf{p}(\mathbf{I} - \mathbf{P}) = \mathbf{0}$, which furnishes $M - 1$ independent linear equations. Then with help of the additional equation $\mathbf{p}\mathbf{1} = 1$, where $\mathbf{1}$ is a size M column

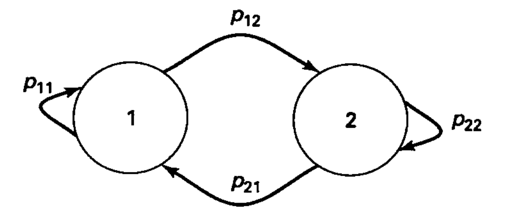


Figure 8.5-1 The state-transition diagram of a two-state Markov chain.

vector of all ones, we can solve for the M values in $\mathbf{p}$. The existence of a steady state, or equivalently asymptotic stationarity, will depend on the eigenvalues of the state-transition matrix $\mathbf{P}$.

Example 8.5-5

(*asymmetric two-state Markov chain*) Here we consider an example of a two-state, asymmetric Markov chain (AMC), with state labels 0 and 1, and state-transition matrix,

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}.$$

See Figure 8.5-2.

Note that in this model there is no requirement that $p_{00} = p_{11}$ and the steady-state probabilities, if they exist, are given by the solution of

$$\mathbf{p}[n+1] = \mathbf{p}[n]\mathbf{P}, \tag{8.5-5}$$

if we let $n \to \infty$. Denoting these probabilities by $p_0[\infty]$ and $p_1[\infty]$, and using $p_0[\infty] + p_1[\infty] = 1$, we obtain

$$p_0[\infty] = \frac{1 - p_{11}}{2 - p_{00} - p_{11}},$$

$$p_1[\infty] = \frac{1 - p_{00}}{2 - p_{00} - p_{11}},$$

which, using the $\mathbf{P}$ matrix from Example 8.5-4, yields $p_0[\infty] = \frac{2}{3}$ and $p_1[\infty] = \frac{1}{3}$.

The *steady-state* autocorrelation function of the AMC of this example can be computed from the Markov state probabilities. For example, assuming asymptotic stationarity,

$$\begin{aligned} R_{XX}[m] &\approx P\{X[k] = 1, X[m+k] = 1\} \quad \text{for sufficiently large } k \\ &= P\{X[k] = 1\}P\{X[m+k] = 1|X[k] = 1\} \\ &= p_1[\infty]\ P\{X[m] = 1|X[0] = 1\}, \end{aligned} \tag{8.5-6}$$

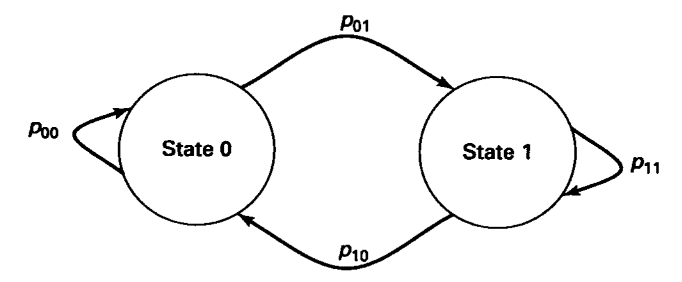


Figure 8.5-2 State-transition diagram of general (asymmetric) two-state Markov random sequence, with state labels 0 and 1.

where the last factor is an m-step transition *from* state 1 *to* state 1. It can be computed recursively from Equation 8.5-5, with the initial condition $\mathbf{p}[0] = [0, 1]$. The needed computation can also be illustrated in a trellis diagram as seen in the following example.

Example 8.5-6

(*trellis diagram for Markov chain*) Consider once again Example 8.1-15, where we introduced the state-transition diagram for what we now know as a Markov chain. Another useful diagram that shows allowable paths to reach a certain state, and the probability of those paths, is the *trellis diagram*, named for its resemblance to the common wooden garden trellis that supports some plants. See Figure 8.5-3 for the two-state case having labels 0 and 1, which also assumes symmetry, that is, $p_{ij} = p_{ji}$. We see that this trellis is a collapsing of the more general tree diagram of Example 8.1-4. The collapse of the tree to the trellis is permitted because of the Markov condition on the conditional probabilities, that serve as the branch labels.

Each node represents the state at a given time instant. The node value (label) is its probability at time n. The links (directed branches) denote possible transitions and are labeled with their respective transition probabilities. Paths through the trellis then represent allowable multiple time-step transitions, with probability given as the product of the transition probabilities along the path.

If we know that the chain is in state one at time $n = 0$, then the *modified* trellis diagram simplifies to that of Figure 8.5-4, where we have labeled the state 1 nodes with

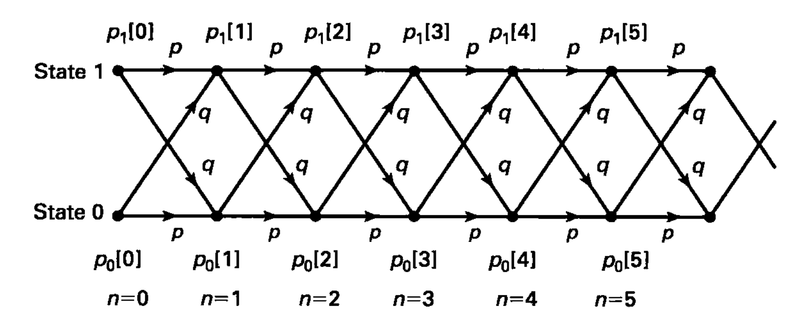


Figure 8.5-3 A trellis diagram of a two-state symmetric Markov chain with state labels 0 and 1. Here $p_i[n]$ is the probability of being in state i at time n.

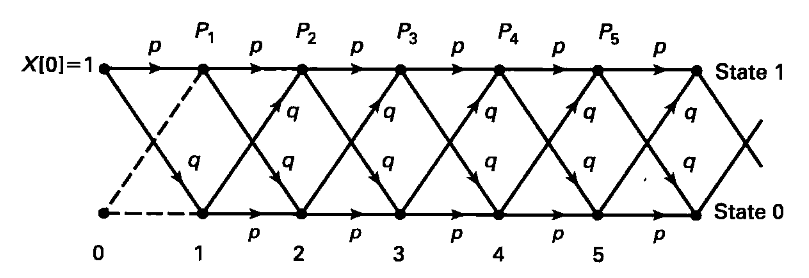


Figure 8.5-4 Trellis diagram conditioned on $X[0] = 1$.

$P_n \triangleq P\{X[n] = 1|X[0] = 1\}$, and we can use this trellis to calculate the probabilities $P\{X[n] = 1|X[0] = 1\}$ needed in Equation 8.5-6. The first few P_n values are easily calculated as $P_1 = p, P_2 = p^2 + q^2, P_3 = p^3 + 3pq^2$, etc. For the case $p_0[\infty] = p_1[\infty] = \frac{1}{2}$, and $p = 0.8$, the *asymptotically stationary autocorrelation* (ASA) function $R_{XX}[m]$ then becomes $R_{XX}[0] = 0.5, R_{XX}[\pm 1] = 0.4, R_{XX}[\pm 2] = 0.34, R_{XX}[\pm 3] = 0.304$, and so forth.†

The trellis diagram shows that, except in trivial cases, there are many allowable paths to reach a certain node, that is, a given state at a given time. This raises the question of which path is most probable (most likely) to make the required multistep traversal. In the previous example, and with $p > q$, it is just a matter of finding the path with the most p's. In general, however, finding the most likely path is a time-consuming problem and, if left to trial-and-error techniques, would quickly exhaust the capabilities of most computers. Much research has been done on this problem because of its many engineering applications, one being speech recognition by computer. In Chapter 11, we discuss the efficient *Viterbi algorithm* for finding the most likely path.

Example 8.5-7

(*buffer fullness*) Consider the Markov chain as a model for a communications buffer with $M+1$ states, with labels 0 to M indicating buffer fullness. In other words, the state label is the number of bytes currently stored in the M byte capacity buffer. Assume that transitions can occur only between neighboring states; that is, the fullness can change at most by one byte in each time unit. The state-transition diagram then appears as shown in Figure 8.5-5.

If we let M go to infinity in Example 8.5-7, we have what is called the general *birth–death* chain, which was first used to model the size of a population over time. In each time unit, there can be at most one birth and at most one death.

Solving the equations. Consider a two-state Markov chain with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}.$$

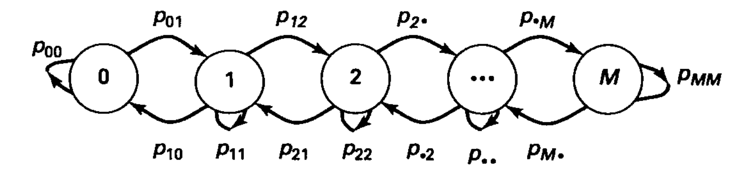


Figure 8.5-5 Markov chain model for $M+1$ state communications buffer.

---

†The ASA is computed as $R_{XX}[m] = E\{X[k+m]X[k]\}$, where $k \to \infty$. For levels of 0 and 1, $R_{XX}[m] = P\{X[m+k] = 1|X[k] = 1\} \times 0.5$. Then clearly $R_{XX}[0] = 1 \times 0.5 = 0.5$, $R_{XX}[1] = 0.8 \times 0.5 = 0.4$, $R_{XX}[2] = [(0.8)^2 + (0.2)^2] \times 0.5 = 0.34$, etc.

We can write the equation relating $\mathbf{p}[n]$ and $\mathbf{p}[n+1]$ then as follows:

$$[p_0[n+1],\ \ p_1[n+1]] = [p_0[n],\ \ p_1[n]] \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}. \tag{8.5-7}$$

This vector equation is equivalent to two scalar difference equations which we have to solve together, that is, two *simultaneous* difference equations. We try a solution of the form

$$p_0[n] = C_0 z^n, \qquad p_1[n] = C_1 z^n.$$

Inserting this attempted solution into Equation 8.5-7 and canceling the common term z^n, we obtain

$$C_0 z = C_0 p_{00} + C_1 p_{10},$$
$$C_1 z = C_0 p_{01} + C_1 p_{11},$$

which implies the following necessary conditions:

$$C_1 = C_0 \left(\frac{z - p_{00}}{p_{10}}\right) = C_0 \left(\frac{p_{01}}{z - p_{11}}\right).$$

This gives a constraint relation between the constants C_0 and C_1 as well as a necessary condition on z, the latter being called the *characteristic equation*

$$(z - p_{00})(z - p_{11}) - p_{10}p_{01} = 0.$$

It turns out that the characteristic equation (CE) can be written, using the determinant function, as

$$\det(z\mathbf{I} - \mathbf{P}) = 0.$$

Solving our two-state equation, we obtain just two solutions z_1 and z_2, one of which must equal 1. (Can you see this? Note that $1 - p_{00} = p_{01}$.) The solutions we have obtained thus far can be written as

$$p_0[n] = C_0 z_i^n, p_1[n] = C_0 \left(\frac{z_i - p_{00}}{p_{10}}\right) z_i^n, i = 1, 2.$$

Since the vector difference equation is linear, we can add the two solutions corresponding to the different values of z_i, to get the general solution, written in row vector form

$$\mathbf{p}[n] = A_1 \left[1, \frac{z_1 - p_{00}}{p_{10}}\right] z_1^n + A_2 \left[1, \frac{z_2 - p_{00}}{p_{10}}\right] z_2^n,$$

where we have introduced two new constants A_1 and A_2 for each of the two linearly independent solutions. These two constants must be evaluated from the initial probability vector $\mathbf{p}[0]$ and the necessary conditions on the probability row vector at time index n, that is, $\sum_{i=0}^{1} p_i[n] = 1$ for all $n \geq 0$.

Example 8.5-8

(*complete solution*) Let

$$\mathbf{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}, \text{ with } \mathbf{p}[0] = \left[1/2, \;\; 1/2\right],$$

and solve for the complete solution, including the startup transient and the steady-state values for $\mathbf{p}[n]$.

The first step is to find the eigenvalues of $\mathbf{P}$, which are the roots of the characteristic equation (CE)

$$\det(z\mathbf{I} - \mathbf{P}) = \det\begin{pmatrix} z - 0.9 & -0.1 \\ -0.2 & z - 0.8 \end{pmatrix} = z^2 - 1.7z + 0.7 = 0.$$

This gives roots $z_1 = 0.7$ and $z_2 = 1.0$. Thus, we can write

$$\mathbf{p}[n] = C_1[1, -1]\, 0.7^n + C_2[1, 0.5]\, 1^n.$$

From steady-state requirement that the components of $\mathbf{p}$ sum to 1.0, we get $C_2 = \frac{2}{3}$. So we can further write

$$\mathbf{p}[n] = C_1[1, -1]\, 0.7^n + \left[\tfrac{2}{3}, \tfrac{1}{3}\right].$$

Finally we invoke the specified initial conditions $\mathbf{p}[0] = [1/2, 1/2]$ to obtain $C_1 = -\frac{1}{6}$ and

$$\mathbf{p}[n] = [-\tfrac{1}{6}, \tfrac{1}{6}]\, 0.7^n + [\tfrac{2}{3}, \tfrac{1}{3}], \text{ or in scalar form,}$$

$$\left.\begin{aligned} p_0[n] &= -\tfrac{1}{6}\, 0.7^n + \tfrac{2}{3} \\ p_1[n] &= \tfrac{1}{6}\, 0.7^n + \tfrac{1}{3} \end{aligned}\right. \quad \text{for } n \geq 0.$$

Here we see that the steady-state probabilities exist and are $p_0[\infty] = \frac{2}{3}$ and $p_1[\infty] = \frac{1}{3}$. The next example shows that such steady-state probabilities do not always exist.

Example 8.5-9

(*ping pong*) Consider the two-state Markov chain with transition probability matrix $\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. The characteristic equation becomes

$$\det(z\mathbf{I} - \mathbf{P}) = \det\begin{pmatrix} z & -1 \\ -1 & z \end{pmatrix} = z^2 - 1 = 0,$$

with two roots $z_{1,2} = \pm 1$. Thus there is no steady state in this case, even though one of the eigenvalues of $\mathbf{P}$ is 1. Indeed, direct from the state-transition diagram, we can see that the random sequence will forever cycle back and forth between states 0 and 1 with each successive time tick. The phase can be irrevocably set by the initial probability vector $\mathbf{p}[0] = [1, 0]$.

While we cannot always assume a steady state exists, note that this example is degenerate in that the transition probabilities into and out of states are either 0 or 1. Another problem for existence of the steady state is a so-called *trapping state*. This is a state with transitions into, but not out of, itself. In most cases of interest in communications and signal processing, a steady state will exist, independent of where the chain starts.

8.6 VECTOR RANDOM SEQUENCES AND STATE EQUATIONS

The scalar random sequence concepts we have seen thus far can be extended to vector random sequences. They are used in Chapter 11 to derive linear estimators for signals in noise (Kalman filter). They are also used in models of sensor arrays, for example, seismic, acoustic, and radar. This section will introduce difference equations for random vectors and the concept of vector Markov random sequence. Interestingly, a high-order Markov-p scalar random sequence can be represented as a first-order vector Markov sequence.

Definition 8.6-1 A *vector random sequence* is a mapping from a probability sample space Ω, corresponding to probability space $(\Omega, \mathscr{F}, P)$, into the space of vector-valued sequences over complex numbers. ■

Thus for each $\zeta \in \Omega$ and fixed time n, we generate a vector $\mathbf{X}(n, \zeta)$. The vector random sequence is usually written $\mathbf{X}[n]$, suppressing the outcome ζ.

For example the first-order CDF for a random vector sequence $\mathbf{X}[n]$, would be given as

$$F_{\mathbf{X}}(\mathbf{x}; n) \triangleq P\{\mathbf{X}[n] \leq \mathbf{x}\},$$

where $\{\mathbf{X}[n] \leq \mathbf{x}\}$ means every element satisfies the inequality, that is, $\{X_1[n] \leq x_1, X_2[n] \leq x_2, \ldots, X_N[n] \leq x_N\}$. Second- and higher-order probabilities would be specified accordingly. The vector random sequence is said to be *statistically specified* by the set of all its first- and higher-order CDFs (or pdf's or PMFs).

The following example treats the correlation analysis for a vector random sequence input to a vector LCCDE

$$\mathbf{y}[n] = \mathbf{A}\mathbf{y}[n-1] + \mathbf{B}\mathbf{x}[n],$$

with N-dimensional coefficient matrices $\mathbf{A}$ and $\mathbf{B}$. In this vector case, BIBO stability is assured when the eigenvalues of $\mathbf{A}$ are less than one in magnitude.

Example 8.6-1

(*vector LCCDEs*) In the vector case, the scalar first-order LCCDE model, excited by column vector random sequence $\mathbf{X}[n]$, becomes

$$\mathbf{Y}[n] = \mathbf{A}\mathbf{Y}[n-1] + \mathbf{B}\mathbf{X}[n], \tag{8.6-1}$$

which is a first-order vector difference equation in the sample vector sequences. The vector impulse response is the column vector sequence

$$\mathbf{h}[n] = \mathbf{A}^n \mathbf{B} u[n],$$

and the zero initial-condition response to the sequence $\mathbf{X}[n]$ is

$$\mathbf{Y}[n] = \sum_{k=0}^{n} \mathbf{A}^{n-k} \mathbf{B} \mathbf{X}[k]$$
$$\triangleq \mathbf{h}[n] * \mathbf{X}[n].$$

The matrix system function is

$$\mathbf{H}(z) = (\mathbf{I} - \mathbf{A} z^{-1})^{-1} \mathbf{B},$$

as can easily be verified. The WSS cross-correlation matrices $\mathbf{R}_{\mathbf{YX}}[m] \triangleq E\{\mathbf{Y}[n+m]\mathbf{X}^{\dagger}[n]\}$ (where the "†" indicates the Hermitian (or conjugate) transpose) and $\mathbf{R}_{\mathbf{XY}}[m] \triangleq E\{\mathbf{X}[n+m]\mathbf{Y}^{\dagger}[n]\}$, between an input WSS random vector sequence and its WSS output random sequence, become

$$\mathbf{R}_{\mathbf{YX}}[m] = \mathbf{h}[m] * \mathbf{R}_{\mathbf{XX}}[m],$$
$$\mathbf{R}_{\mathbf{XY}}[m] = \mathbf{R}_{\mathbf{XX}}[m] * \mathbf{h}^{\dagger}[-m].$$

Parenthetically, we note that for a causal $\mathbf{h}$, such as would arise from recursive solution of the above vector LCCDE, we have the output $\mathbf{Y}[n]$ uncorrelated with the future values of the input $\mathbf{X}[n]$, when the input $\mathbf{X} = \mathbf{W}$ is assumed a white noise vector sequence.

The output correlation matrix is

$$\mathbf{R}_{\mathbf{YY}}[m] = \mathbf{h}[m] * \mathbf{R}_{\mathbf{XX}}[m] * \mathbf{h}^{\dagger}[-m]$$

and the output psd matrix becomes upon Fourier transformation

$$\mathbf{S}_{\mathbf{YY}}(\omega) = \mathbf{H}(\omega)\mathbf{S}_{\mathbf{XX}}(\omega)\mathbf{H}^{\dagger}(\omega).$$

The total solution of Equation 8.6-1 for any $n \geq n_0$ can be written as

$$\mathbf{Y}[n] = \mathbf{A}^{n-n_0}\mathbf{Y}[n_0] + \sum_{k=n_0}^{n} \mathbf{h}[n-k]\mathbf{X}[k], \qquad n \geq n_0$$

in terms of the initial condition $\mathbf{Y}[n_0]$ that must be specified at n_0. In the limit as $n_0 \to -\infty$, and for a stable system matrix $\mathbf{A}$, this then becomes the convolution summation

$$\mathbf{Y}[n] = \mathbf{h}[n] * \mathbf{X}[n], \qquad -\infty < n < +\infty.$$

Definition 8.6-2 A vector random sequence $\mathbf{Y}[n]$ is *vector Markov* if for all $K > 0$ and for all $n_K > n_{K-1} > \ldots > n_1$, we have

$$P\{\mathbf{Y}[n_K] \leq \mathbf{y}_K | \mathbf{y}[n_{K-1}], \ldots, \mathbf{y}[n_1]\} = P\{\mathbf{Y}[n_K] \leq \mathbf{y}_K | \mathbf{y}[n_{K-1}]\}$$

for all real values of the vector $\mathbf{y}_K$, and all conditioning vectors $\mathbf{y}[n_{K-1}], \ldots, \mathbf{y}[n_1]$. (cf. Definition 8.5-2 of Markov-p property.) ■

We can now state the following theorem for vector random sequences:

Theorem 8.6-1 In the state equation

$$\mathbf{X}[n] = \mathbf{A}\mathbf{X}[n-1] + \mathbf{B}\mathbf{W}[n], \text{for } n > 0, \text{ with } \mathbf{X}[0] = \mathbf{0},$$

let the input $\mathbf{W}[n]$ be a white Gaussian random sequence. Then the output $\mathbf{X}[n]$ for $n > 0$ is a vector Markov random sequence.

The proof is left to the reader as an exercise. ■

Example 8.6-2

(*relation between scalar Markov-p and vector Markov*) Let $X[n]$ be a Markov-p random sequence satisfying the pth order difference equation

$$X[n] = a_1 X[n-1] + \ldots + a_p X[n-p] + bW[n].$$

Defining the p-dimensional vector random sequence $\mathbf{X}[n] = [X[n], \ldots, X[n-p+1]]^T$, and coefficient matrix

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 & \cdots & \cdots & a_p \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix},$$

we have

$$\mathbf{X}[n] = \mathbf{A}\mathbf{X}[n-1] + \mathbf{b}W[n].$$

Thus $\mathbf{X}[n]$ is a vector Markov random sequence with $\mathbf{b} = [b, 0, \ldots, 0]^T$. Such a vector transformation of a scalar equation is called a *state-variable representation* [8-7].

8.7 CONVERGENCE OF RANDOM SEQUENCES

Some nonstationary random sequences may converge to a limit as the sequence index goes to infinity, for example as time becomes infinite. This asymptotic behavior is evidenced in probability theory by convergence of the fraction of successes in an infinite Bernoulli sequence, where the relevant theorems are called the laws of large numbers. Also, when we study the convergence of random processes in Chapter 10 we will sometimes make a sequence of finer and finer approximations to the output of a random system at a given time, say t_0, that is, $Y_n(t_0)$. The index n then defines a random sequence, which should converge in some sense to the true output. In this section we will look at several types of convergence for random sequences, that is, sequences of random variables.

We start by reviewing the concept of convergence for deterministic sequences. Let x_n be a sequence of complex (or real) numbers; then convergence is defined as follows.

Definition 8.7-1 A sequence of complex (or real) numbers x_n converges to the complex (or real) number x if given any $\varepsilon > 0$, there exists an integer n_0 such that whenever $n > n_0$, we have

$$|x_n - x| < \varepsilon. \quad \blacksquare$$

Note that in this definition the value n_0 may depend on the value ε; that is, when ε is made smaller, most likely n_0 will need to be made larger. Sometimes this dependence is formalized by writing $n_0(\varepsilon)$ in place of n_0 in this definition. This is often written as

$$\lim_{n\to\infty} x_n = x \qquad \text{or as} \qquad x_n \to x \text{ as } n \to \infty.$$

A practical problem with this definition is that one must have the limit x to test for convergence. For simple cases one can often guess what the limit is and then use the definition to verify that this limit indeed exists. Fortunately, for more complex situations there is an alternative in the *Cauchy criterion* for convergence, which we state as a theorem without proof.

Theorem 8.7-1 (*Cauchy criterion* [8-8]) A sequence of complex (or real) numbers x_n converges to a limit *if and only if* (iff)

$$|x_n - x_m| \to 0 \text{ as both } n \text{ and } m \to \infty.$$

The reason that this works for complex (or real) numbers is that the set of all complex (or real) numbers is *complete*, meaning that it contains all its limit points. For example, the set $\{0 < x < 1\} = (0, 1)$ is not complete, but the set $\{0 \leq x \leq 1\} = [0, 1]$ is complete because sequences x_n in these sets and tending to 0 or 1 have a *limit point* in the set $[0, 1]$ but have no limit point in the set $(0, 1)$. In fact, the set of all complex (or real) numbers is complete as well as n-dimensional linear vector spaces over both the real and complex number fields. Thus the Cauchy criterion for convergence applies in these cases also. For more on numerical convergence see [8-8].

Convergence for functions is defined using the concept of convergence of sequences of numbers. We say the sequence of functions $f_n(x)$ converges to the function $f(x)$ if the corresponding sequence of numbers converges for each x. It is stated more formally in the following definition.

Definition 8.7-2 The sequence of functions $f_n(x)$ converges (pointwise) to the function $f(x)$ if for each x_0 the sequence of complex numbers $f_n(x_0)$ converges to $f(x_0)$. $\blacksquare$

The Cauchy criterion for convergence applies to pointwise convergence of functions if the set of functions under consideration is complete. The set of continuous functions is not complete because a sequence of continuous functions may converge to a discontinuous function (cf. item (d) in Example 8.7-1). However, the set of bounded functions is complete [8-8].

The following are some examples of convergent sequences of numbers and functions. We leave the demonstration of these results as exercises for the reader.

Example 8.7-1
(*some convergent sequences*)

(a) $x_n = (1 - 1/n)a + (1/n)b \rightarrow a$ as $n \rightarrow \infty$.
(b) $x_n = \sin(\omega + e^{-n}) \rightarrow \sin\omega$ as $n \rightarrow \infty$.
(c) $f_n(x) = \sin[(\omega + 1/n)x] \rightarrow \sin(\omega x)$, as $n \rightarrow \infty$ for any (fixed) x.
(d) $f_n(x) = \left\{ \begin{array}{ll} e^{-n^2 x}, & \text{for } x > 0 \\ 1, & \text{for } x \leq 0 \end{array} \right\} \rightarrow u(-x)$, as $n \rightarrow \infty$ for any (fixed) x.

The reader should note that in the convergence of the functions in (c) and (d), the variable x is held constant as the limit is being taken. The limit is then repeated for each such x value to find the limiting function.

Since a random variable is a function, a sequence of random variables (also called a random sequence) is a sequence of functions. Thus, we can define the first and strongest type of convergence for random variables.

Definition 8.7-3 (*sure convergence*) The random sequence $X[n]$ *converges surely* to the random variable X if the sequence of functions $X[n, \zeta]$ converges to the function $X(\zeta)$ as $n \rightarrow \infty$ for all outcomes $\zeta \in \Omega$. ■

As a reminder, the functions $X(\zeta)$ are not arbitrary. They are random variables and thus satisfy the condition that the set $\{\zeta\colon X(\zeta) \leq x\} \subset \mathscr{F}$ for all x, that is, that this set be an event for all values of x. This is in fact necessary for the calculation of probability since the probability measure P is defined only for events. Such functions X are more generally called *measurable functions* and in a course on real analysis it is shown that the space of measurable functions is complete [8-1]. If we have a Cauchy sequence of measurable functions (random variables), then one can show that the limit function exists and is also measurable (a random variable). Thus, the Cauchy convergence criterion also applies for random variables.

Most of the time we are not interested in precisely defining random variables for sets in Ω of probability zero because it is thought that these events will never occur. In this case, we can weaken the concept of sure convergence to the still very strong concept of almost-sure convergence.

Definition 8.7-4 (*almost-sure convergence*) The random sequence $X[n]$ converges *almost surely* to the random variable X if the sequence of functions $X[n, \zeta]$ converges for all outcomes $\zeta \in \Omega$ except possibly on a set of probability zero. ■

This is the strongest type of convergence normally used in probability theory. It is also called *probability-1* convergence. It is sometimes written

$$P\left\{\lim_{n\rightarrow\infty} X[n, \zeta] = X(\zeta)\right\} = 1,$$

meaning simply that there is a set A such that $P[A] = 1$ and $X[n]$ converges to X for all $\zeta \in A$. In particular $A \triangleq \{\zeta\colon \lim_{n\rightarrow\infty} X[n, \zeta] = X(\zeta)\}$. Here the set A^c is the probability-zero set mentioned in this definition. As shorthand notation we also use

$$X[n] \to X \text{ a.s.} \quad \text{and} \quad X[n] \to X \quad \text{pr.1},$$

where the abbreviation "a.s." stands for *almost surely*, and "pr.1" stands for *probability 1.*

An example of probability-1 convergence is the Strong Law of Large Numbers to be proved in the next section. Three examples of random sequences are next evaluated for possible convergence.

Example 8.7-2

(*convergence of random sequences*) For each of the following three random sequences, we assume that the probability space $(\Omega, \mathscr{F}, P)$ has sample space $\Omega = [0,1]$. $\mathscr{F}$ is the family of Borel subsets of Ω and the probability measure P is Lebesgue measure, which on a real interval $(a,b]$ is just its length l, that is,

$$l(a,b] \triangleq b - a \quad \text{for } b \geq a.$$

(a) $X[n,\zeta] = n\zeta$.
(b) $X[n,\zeta] = \sin(n\zeta)$.
(c) $X[n,\zeta] = \exp[-n^2(\zeta - \frac{1}{n})]$.

The sequence in (a) clearly diverges to $+\infty$ for any $\zeta \neq 0$. Thus this random sequence does not converge. The sequence in (b) does not diverge, but it oscillates between -1 and $+1$ except for the one point $\zeta = 0$. Thus this random sequence does not converge either.

Considering the random sequence in (c), the graph in Figure 8.7-1 shows that this sequence converges as follows:

$$\lim_{n\to\infty} X[n,\zeta] = \begin{cases} \infty & \text{for } \zeta = 0 \\ 0 & \text{for } \zeta > 0. \end{cases}$$

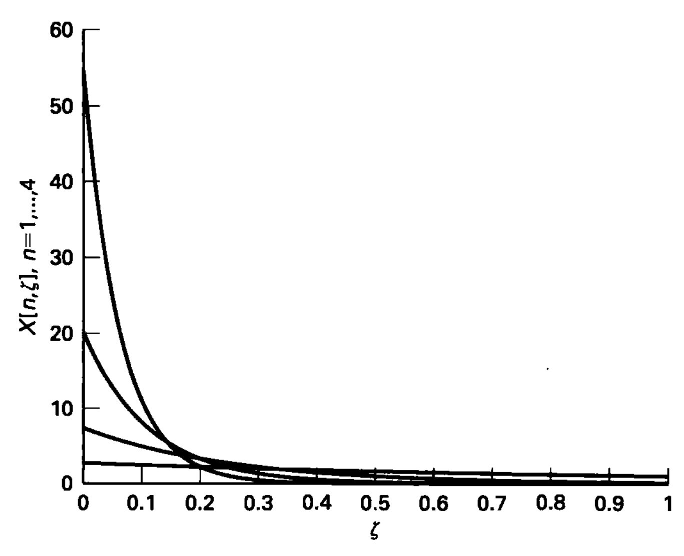


Figure 8.7-1 Plot of sequence (c) $X[n,\zeta]$ versus ζ for $\Omega = [0,1]$ for $n = 1,\ldots,4$.

Thus, we can say that the random sequence converges to the (degenerate) random variable $X = 0$ with probability 1. We simply take $A = (0, 1]$ and note that $P[A] = 1$ and that $X[n, \zeta] \simeq 0$ for every ζ in A for sufficiently large n. We write $X[n] \to 0$ a.s. However, $X[n]$ clearly does not converge surely to zero.

Thus far we have been discussing pointwise convergence of sequences of functions and random sequences. This is similar to considering a space of *bounded* functions $\mathscr{B}$ with the norm

$$||f||_\infty \triangleq \sup_x |f(x)|.^\dagger$$

When we write $f_n \to f$ in the function space $\mathscr{B}$, we mean that $||f_n - f||_\infty = \sup_x |f_n(x) - f(x)| \to 0$, giving us pointwise convergence. The space of continuous bounded functions is denoted L_∞ and is known to be complete ([8-1], p. 115).

Another type of function space of great practical interest uses the *energy norm* (cf. Equation 4.4-6):

$$||f||_2 \triangleq \left(\int_{-\infty}^{+\infty} |f(x)|^2 dx \right)^{1/2}.$$

The space of integrable (measurable) functions with *finite energy norm* is denoted L^2. When we say a sequence of functions converges in L^2, that is, $||f_n - f||_2 \to 0$, we mean that

$$\left(\int_{-\infty}^{+\infty} |f_n(x) - f(x)|^2 dx \right)^{1/2} \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

This space of integrable functions is also complete [8-1]. A corresponding concept for random sequences is given by mean-square convergence.

Definition 8.7-5 (mean-square convergence) A random sequence $X[n]$ converges in the *mean-square sense* to the random variable X if $E\{|X[n] - X|^2\} \to 0$ as $n \to \infty$. ■

This type of convergence depends only on the second-order properties of the random variables and is thus often easier to calculate than a.s. convergence. A second benefit of the mean-square type of convergence is that it is closely related to the physical concept of power. If $X[n]$ converges to X in the mean-square sense, then we can expect that the variance of the error $\varepsilon[n] \triangleq X[n] - X$ will be small for large n. If we look back at Example 8.7-2, (c), we can see that this random sequence does not converge in the mean-square sense, so that the error variance or power as defined here would not ever be expected to be small. To see this, consider possible mean-square convergence to zero (since $X[n] \to 0$ a.s.),

---

†The supremum or sup operator is similar to the max operator. The supremum of a set of numbers is the smallest number greater than or equal to each number in the set, for example, $\sup\{0 < x < 1\} = 1$. Note the difficulty with max in this example since 1 is not included in the open interval $(0, 1)$; thus the max does not exist here!

$$
\begin{aligned}
E\{|X[n]-0|^2\} &= E\{X[n]^2\} \\
&= \int_0^1 \exp(-2n^2\zeta)\exp 2n d\zeta \\
&= \exp(2n)\int_0^1 \exp(-2n^2\zeta)d\zeta \\
&= \exp(2n)\left[\frac{1-\exp(-2n^2)}{2n^2}\right] \to \infty \quad \text{as } n \to \infty.
\end{aligned}
$$

Hence $X[n]$ does not converge in the mean-square sense to 0.

Still another type of convergence that we will consider is called *convergence in probability*. It is weaker than probability-1 convergence and also weaker than mean-square convergence. This is the type of convergence displayed in the Weak Laws of Large Numbers to be discussed in the next section. It is defined as follows:

Definition 8.7-6 (*convergence in probability*) Given the random sequence $X[n]$ and the limiting random variable X, we say that $X[n]$ converges in probability to X if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left[|X[n]-X| > \varepsilon\right] = 0. \quad \blacksquare$$

We sometimes write $X[n] \to X(p)$, where (p) denotes the type of convergence. Also convergence in probability is sometimes called p-convergence.

One can use Chebyshev's inequality (Theorem 4.4-1), $P[|Y| > \varepsilon] \leq E[|Y|^2]/\varepsilon^2$ for $\varepsilon > 0$, to show that mean-square convergence implies convergence in probability. For example, let $Y \triangleq X[n] - X$; then the preceding inequality becomes

$$P[|X[n]-X| > \varepsilon] \leq E\left[|X[n]-X|^2\right]/\varepsilon^2.$$

Now mean-square convergence implies that the right-hand side goes to zero as $n \to \infty$, for any fixed $\varepsilon > 0$, which implies that the left-hand side must also go to zero, which is the definition of convergence in probability. Thus we have proved the following result.

Theorem 8.7-2 Convergence of a random sequence in the mean-square sense implies convergence in probability. $\blacksquare$

The relation between convergence with probability 1 and convergence in probability is more subtle. The main difference between them can be seen by noting that the former talks about the probability of the limit while the latter talks about the limit of the probability. Further insight can be gained by noting that a.s. convergence is concerned with convergence of the entire sample sequences while p-convergence is concerned only with the convergence of the random variable at an individual n. That is to say, a.s. convergence is concerned with the joint events at an infinite number of times, while p-convergence is concerned with the simple event at time n, albeit large. One can prove the following theorem.

Theorem 8.7-3 Convergence with probability 1 implies convergence in probability.

Proof (adapted from Gnedenko [8-9].) Let $X[n] \to X$ a.s. and define the set A,

$$A \triangleq \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{m=1}^{\infty} \{\zeta\colon |X[n+m,\zeta] - X(\zeta)| < 1/k\}.$$

Then it must be that $P[A] = 1$. To see this we note that A is the set of ζ such that starting at some n and for all later n we have $|X[n,\zeta] - X(\zeta)| < 1/k$ and furthermore this must hold for all $k > 0$. Thus, A is precisely the set of ζ on which $X[n,\zeta]$ is convergent. So $P[A]$ must be 1. Eventually for n large enough and $1/k$ small enough we get $|X[n,\zeta] - X(\zeta)| < \varepsilon$, and the error stays this small for all larger n. Thus,

$$P\left[\bigcup_{n=1}^{\infty} \bigcap_{m=1}^{\infty} \{|X[n+m] - X| < \varepsilon\}\right] = 1 \qquad \text{for all } \varepsilon > 0,$$

which implies by the continuity of probability,

$$\lim_{n\to\infty} P\left[\bigcap_{m=1}^{\infty} \{|X[n+m] - X| < \varepsilon\}\right] = 1 \qquad \text{for all } \varepsilon > 0,$$

which in turn implies the greatly weakened result

$$\lim_{n\to\infty} P[|X[n+m] - X| < \varepsilon] = 1 \qquad \text{for all } \varepsilon > 0, \tag{8.7-1}$$

which is equivalent to the definition of p-convergence. ■

Because of the gross weakening of the a.s. condition, that is, the enlargement of the set A in the foregoing proof, it can be seen that p-convergence does not imply a.s. convergence. We note in particular that Equation 8.7-1 may well be true even though no single sample sequence stays close to X for all $n+m > n$. This is in fact the key difference between these two types of convergence.

Example 8.7-3

(*a convergent random sequence?*) Define a random pulse sequence $X[n]$ on $n \geq 0$ as follows: Set $X[0] = 1$. Then for the *next two points* set exactly one of the $X[n]$'s to 1, equally likely among the two points, and the other to 0. For the *next three points* set exactly one of the $X[n]$'s to 1 equally likely among the three points and set the others to 0. Continue this procedure for the *next four points*, setting exactly one of the $X[n]$'s to 1 equally likely among the four points and the others to 0 and so forth. A sample function would look like Figure 8.7-2.

Obviously this random sequence is slowly converging to zero in some sense as $n \to \infty$. In fact a simple calculation would show p-convergence and also mean-square convergence due to the growing distance between pulses as $n \to \infty$. In fact at $n \simeq \frac{1}{2}l^2$, the probability of a one (pulse) is only $1/l$. However, we do not have a.s. convergence, since *every* sample sequence has ones appearing arbitrarily far out on the n axis. Thus no sample sequences converge to zero.

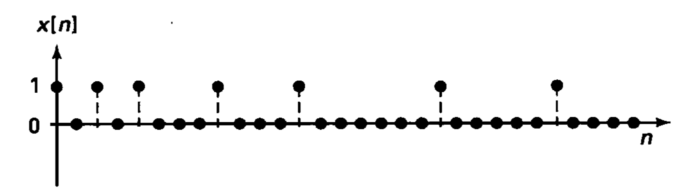


Figure 8.7-2 A sequence that is converging in probability but not with probability 1.

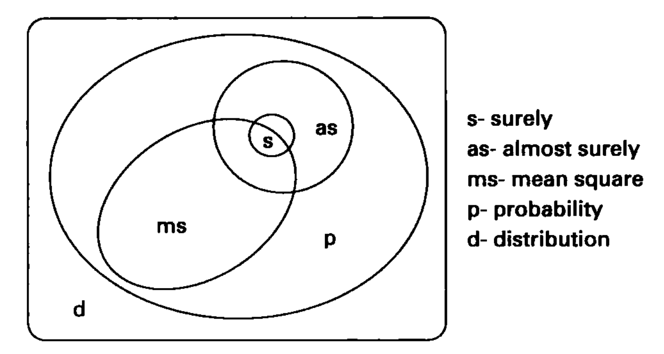


Figure 8.7-3 Venn diagram illustrating relationships of various possible convergence modes for random sequences.

One final type of convergence that we consider is not a convergence for random variables at all! Rather it is a type of convergence for distribution functions.

Definition 8.7-7 A random sequence $X[n]$ with CDF $F_n(x)$ converges in distribution to the random variable X with CDF $F(x)$ if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at all x for which F is continuous. ■

Note that in this definition we are not really saying anything about the random variables themselves, just their CDFs. Convergence in distribution just means that as n gets large the CDFs are converging or becoming alike. For example, the sequence $X[n]$ and the variable X can be jointly independent even though $X[n]$ converges to X in distribution. This is radically different from the four earlier types of convergence, where as n gets large the random variables $X[n]$ and X are becoming very dependent because some type of "error" between them is going to zero. Convergence in distribution is the type of convergence that occurs in the Central Limit Theorem (see Section 4.7). The relationships between these five types of convergence are shown diagrammatically in Figure 8.7-3, where we have used the fact that p-convergence implies convergence in distribution, which is shown below. Note that even sure convergence may not imply mean-square convergence. This because the integral of the square of the limiting random variable, with respect to the probability measure, may diverge.

To see that p-convergence implies convergence in distribution, assume that the limiting random variable X is continuous so that it has a pdf. First we consider the conditional distribution function

$$F_{X[n]|X}(y|x) = P\{X[n] \leq y | X = x\}.$$

From the definition of p-convergence, it should be clear that

$$F_{X[n]|X}(y|x) \to \begin{cases} 1, & y > x \\ 0, & y < x \end{cases}, \qquad \text{as } n \to \infty,$$

so that

$$F_{X[n]|X}(y|x) \to u(y-x), \qquad \text{except possibly at the one point } y = x,$$

and hence

$$\begin{aligned} F_{X[n]}(y) = P\{X[n] \leq y\} &= \int_{-\infty}^{+\infty} F_{X[n]|X}(y|x) f_X(x)dx \\ &\to \int_{-\infty}^{+\infty} u(y-x) f_X(x)dx \\ &= \int_{-\infty}^{y} f_X(x)dx \\ &= F_X(y), \end{aligned}$$

as was to be shown. In the case where the limiting random variable X is not continuous, we must exercise more care but the result is still true at all points x for which $F_X(x)$ is continuous. (See Problem 8.54.)

8.8 LAWS OF LARGE NUMBERS

The Laws of Large Numbers have to do with the convergence of a sequence of estimates of the mean of a random variable. As such they concern the convergence of a random sequence to a constant. The Weak Laws obtain convergence in probability, while the Strong Laws yield convergence with probability 1. A version of the Weak Law has already been demonstrated in Example 4.4-3. We restate it here for convenience.

Theorem 8.8-1 (*Weak Law of Large Numbers*) Let $X[n]$ be an independent random sequence with mean μ_X and variance σ_X^2 defined for $n \geq 1$. Define another random sequence as

$$\hat{\mu}_X[n] \triangleq (1/n) \sum_{k=1}^{n} X[k] \qquad \text{for } n \geq 1.$$

Then $\hat{\mu}_X[n] \to \mu_x \quad (p) \quad \text{as } n \to \infty$. ■

Remember, an independent random sequence is one whose terms are all jointly independent. Another version of the Weak Law allows the random sequence to be of nonuniform variance.

Theorem 8.8-2 (*Weak Law—nonuniform variance*) Let $X[n]$ be an independent random sequence with constant mean μ_X and variance $\sigma_X^2[n]$ defined for $n \geq 1$. Then if

$$\sum_{n=1}^{\infty} \sigma_X^2[n]/n^2 < \infty,$$

$$\hat{\mu}_X[n] \rightarrow \mu_X \quad (p) \quad \text{as } n \rightarrow \infty. \quad \blacksquare$$

Both of these theorems are also true for convergence with probability 1, in which case they become Strong Laws. The theorems concerning convergence with probability 1 are best derived using the concept of a Martingale sequence. By introducing this concept we can also get another useful result called the Martingale convergence theorem, which is helpful in estimation and decision/detection theory.

Definition 8.8-1 †A random sequence $X[n]$ defined for $n \geq 0$ is called a *Martingale* if the conditional expectation

$$E\{X[n]|X[n-1], X[n-2], \ldots, X[0]\} = X[n-1] \qquad \text{for all } n \geq 1. \quad \blacksquare$$

Viewing the conditional expectation as an estimate of the present value of the sequence based on the past, then for a Martingale this estimate is just the most recent past value. If we interpret $X[n]$ as an amount of capital in a betting game, then the Martingale condition can be regarded as necessary for fairness of the game, which in fact is how it was first introduced [8-1].

Example 8.8-1

(*binomial counting sequence*) Let $W[n]$ be a Bernoulli random sequence taking values ± 1 with equal probability and defined for $n \geq 0$. Let $X[n]$ be the corresponding Binomial counting sequence

$$X[n] \triangleq \sum_{k=0}^{n} W[k], \qquad n \geq 0.$$

Then $X[n]$ is a Martingale, which can be shown as follows:

$$E\{X[n]|X[n-1], \ldots, X[0]\} = E\left\{\sum_{k=0}^{n} W[k]|X[n-1], \ldots, X[0]\right\}$$

$$= \sum_{k=0}^{n} E\{W[k]|X[n-1], \ldots, X[0]\}$$

†The material dealing with Martingale sequences can be omitted on a first reading.

$$= \sum_{k=0}^{n} E\{W[k]|W[n-1], \ldots, W[0]\}$$

$$= \sum_{k=0}^{n-1} W[k] + E\{W[n]\}$$

$$= X[n-1].$$

The first equality follows from the definition of $X[n]$. The third equality follows from the fact that knowledge of the first $(n-1)$ Xs is equivalent to knowledge of the first $(n-1)$ Ws. The next-to-last equality follows from $E[W|W] = W$. The last equality follows from the fact that $E\{W[n]\} = 0$.

Example 8.8-2

(*independent-increments sequences*) Let $X[n]$ be an independent-increments random sequence (see Definition 8.1-4) defined for $n \geq 0$. Then $X_c[n] \triangleq X[n] - \mu_X[n]$ is a Martingale. To show this we write $X_c[n] = (X_c[n] - X_c[n-1]) + X_c[n-1]$ and note that by independent increments and the fact that the mean of X_c is zero, we have

$$\begin{aligned} E\{X_c[n]|X_c[n-1], \ldots, X_c[0]\} &= E\{X_c[n] - X_c[n-1]|X_c[n-1], \ldots, X_c[0]\} \\ &\quad + E\{X_c[n-1]|X_c[n-1], \ldots, X_c[0]\} \\ &= E\{X_c[n] - X_c[n-1]\} + X_c[n-1] \\ &= X_c[n-1]. \end{aligned}$$

The next theorem shows the connection between the Strong Laws, which have to do with the convergence of sample sequences, and Martingales. It provides a kind of Chebyshev inequality for the maximum term in an n-point Martingale sequence.

Theorem 8.8-3 Let $X[n]$ be a Martingale sequence defined on $n \geq 0$. Then for every $\varepsilon > 0$ and for any positive n,

$$P\left[\max_{0 \leq k \leq n} |X[k]| \geq \varepsilon\right] \leq E\{X^2[n]\}/\varepsilon^2.$$

Proof For $0 \leq j \leq n$, define the mutually exclusive events,

$$A_j \triangleq \{|X[k] \geq \varepsilon \text{ for the } \textit{first time} \text{ at } j\}.$$

Then the event $\{\max_{0 \leq k \leq n} |X[k]| \geq \varepsilon\}$ is just a union of these events. Also define the random variables,

$$I_j \triangleq \begin{cases} 1, & \text{if } A_j \text{ occurs,} \\ 0, & \text{otherwise,} \end{cases}$$

called the *indicators* of the events A_j. Then

$$E\{X^2[n]\} \geq \sum_{j=0}^{n} E\{X^2[n]I_j\} \tag{8.8-1}$$

since $\sum_{j=0}^{n} I_j \leq 1$. Also $X^2[n] = (X[j] + (X[n] - X[j]))^2$, so expanding and inserting into Equation 8.8-1 we get

$$\begin{aligned} E\{X^2[n]\} &\geq \sum_{j=0}^{n} E\{X^2[j]I_j\} + 2\sum_{j=0}^{n} E\{X[j]\,(X[n] - X[j])\,I_j\} \\ &\quad + \sum_{j=0}^{n} E\{(X[n] - X[j])^2\, I_j\} \\ &\geq \sum_{j=0}^{n} E\{X^2[j]I_j\} + 2\sum_{j=0}^{n} E\{X[j]\,(X[n] - X[j])\,I_j\}. \end{aligned} \tag{8.8-2}$$

Letting $Z_j \triangleq X[j]I_j$, we can write the second term in Equation 8.8-2 as $E\{Z_j\,(X[n] - X[j])\}$ and noting that Z_j depends only on $X[0], \ldots, X[j]$, we then have

$$\begin{aligned} E\{Z_j\,(X[n] - X[j])\} &= E\{E[Z_j\,(X[n] - X[j])\,|X[0], \ldots, X[j]]\} \\ &= E\{Z_j E\,[X[n] - X[j]|X[0], \ldots, X[j]]\} \\ &= E\{Z_j\,(X[j] - X[j])\} \\ &= 0. \end{aligned}$$

Thus Equation 8.8-2 becomes

$$\begin{aligned} E\{X^2[n]\} &\geq \sum_{j=0}^{n} E\{X^2[j]I_j\} \\ &\geq \varepsilon^2 E\left\{\sum_{j=0}^{n} I_j\right\} \\ &= \varepsilon^2 P\left\{\bigcup_{j=0}^{n} A_j\right\} \\ &= \varepsilon^2 P\left\{\max_{0\leq k\leq n} |X[k]| \geq \varepsilon\right\}. \quad \blacksquare \end{aligned}$$

Theorem 8.8-4 (*Martingale Convergence theorem*) Let $X[n]$ be a Martingale sequence on $n \geq 0$, satisfying

$$E\{X^2[n]\} \leq C < \infty \quad \text{for all } n \text{ for some } C.$$

Then

$$X[n] \to X \quad \text{(a.s.)} \quad \text{as } n \to \infty,$$

where X is the limiting random variable.

Proof Let $m \geq 0$ and define $Y[n] \triangleq X[n+m] - X[m]$ for $n \geq 0$. Then $Y[n]$ is a Martingale, so by Theorem 8.8-3

$$P\left[\max_{0\leq k\leq n} |X[m+k] - X[m]| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} E\left\{Y^2[n]\right\},$$

where

$$\begin{aligned} E\{Y^2[n]\} &= E\{(X[n+m] - X[m])^2\} \\ &= E\{X^2[n+m]\} - 2E\{X[n+m]X[m]\} + E\{X^2[m]\}. \end{aligned}$$

Rewriting the middle term, we have

$$\begin{aligned} E\{X[m]X[n+m]\} &= E\{X[m]E[X[n+m]|X[m],\ldots,X[0]]\} \\ &= E\{X[m]X[m]\} \\ &= E\{X^2[m]\} \quad \text{since } X \text{ is a Martingale,} \end{aligned}$$

so

$$E\{Y^2[n]\} = E\{X^2[n+m]\} - E\{X^2[m]\} \geq 0 \qquad \text{for all } m, n \geq 0. \tag{8.8-3}$$

Therefore $E\{X^2[n]\}$ must be monotonic nondecreasing. Since it is bounded from above by $C < \infty$, it must converge to a limit. Since it has a limit, then by Equation 8.8-3, the $E\{Y^2[n]\} \to 0$ as m and $n \to \infty$. Thus,

$$\lim_{m\to\infty} P\left[\max_{k\geq 0} |X[m+k] - X[m]| > \varepsilon\right] = 0,$$

which implies $P[\lim_{n\to\infty} \max_{k\geq 0} |X[m+k] - X[m]| > \varepsilon] = 0$ by the continuity of the probability measure P (cf. Corollary to Theorem 8.1-1). Finally by the Cauchy convergence criteria, there exists a random variable X such that

$$X[n] \to X \quad \text{(a.s.)}. \quad \blacksquare$$

Theorem 8.8-5 (*Strong Law of Large Numbers*) Let $X[n]$ be a WSS independent random sequence with mean μ_X and variance σ_X^2 defined for $n \geq 1$. Then as $n \to \infty$

$$\hat{\mu}_X[n] = \frac{1}{n}\sum_{k=1}^{n} X[k] \to \mu_X \quad \text{(a.s.)}.$$

Proof Let $Y[n] \triangleq \sum_{k=1}^{n} \frac{1}{k} X_c[k]$; then $Y[n]$ is a Martingale on $n \geq 1$. Since

$$E\{Y^2[n]\} = \sum_{k=1}^{n} \frac{1}{k^2}\sigma_X^2 \le \sigma_X^2 \sum_{k=1}^{\infty} \frac{1}{k^2} = C < \infty,$$

we can apply Theorem 8.8-4 to show that $Y[n] \to Y$ (a.s.) for some random variable Y. Next noting that $X_c[k] = k\,(Y[k] - Y[k-1])$, we can write

$$\begin{aligned}\frac{1}{n}\sum_{k=1}^{n} X_c[k] &= \frac{1}{n}\left[\sum_{k=1}^{n} kY[k] - \sum_{k=1}^{n} kY[k-1]\right] \\ &= -\frac{1}{n}\sum_{k=1}^{n} Y[k] + \frac{n+1}{n}Y[n] \\ &\to -Y + Y = 0 \qquad \text{(a.s.)}\end{aligned}$$

so that

$$\hat{\mu}_X[n] \to \mu_X \quad \text{(a.s.)}. \quad \blacksquare$$

SUMMARY

In this chapter we introduced the concept of a random sequence and studied its properties and ways to characterize it. We defined the random sequence as a family of sample sequences each associated with an outcome or point in the sample space. We introduced several important random sequences. Then we reviewed linear discrete-time theory and considered the practical problem of finding out how sample sequences are modified as they pass through the system. Our emphasis was on how the mean and covariance function are transformed by a linear system. We then considered the special but important case of stationary and WSS random sequences and introduced the concept of power spectral density for them. We looked at convergence of random sequences and learned to appreciate the variety of modes of convergence that are possible. We then applied some of these results to the laws of large numbers and used Martingale properties to prove the important strong law of large numbers.

Some additional sources for the material in this chapters are [8-9], [8-10], and [8-11].

In the next chapter we will discover that many of these results extend to the case of continuous time as we continue our study with random processes.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

8.1 Consider the following autoregressive process.

$$W_n = 2W_{n-1} + X_n, W_0 = 0$$

$$Z_n = \frac{1}{2}Z_{n-1} + X_n, Z_0 = 0$$

Express W_n and Z_n in terms of $X_n, X_{n-1}, \ldots, X_1$ and then find $E[W_n]$ and $E[Z_n]$.

***8.2** Consider an N-dimensional random vector $\mathbf{X}$. Show that pairwise independence of its random variable components does not imply that the components are jointly independent.

8.3 Let $\mathbf{X} = (X_1, X_2, \ldots, X_5)^T$ be a random vector whose components satisfy the equations

$$X_i = X_{i-1} + B_i, \quad 1 \leq i \leq 5,$$

where the B_i, are jointly independent and Bernoulli distributed, taking on values 0 and 1, with mean value 1/2. The first value is $X_1 = B_1$. Put the B_i together to make a random vector $\mathbf{B}$.

(a) Write $\mathbf{X} = \mathbf{AB}$ for some constant matrix $\mathbf{A}$ and determine $\mathbf{A}$.
(b) Find the mean vector $\boldsymbol{\mu}_{\mathbf{X}}$.
(c) Find the covariance matrix $\mathbf{K}_{\mathbf{BB}}$.
(d) Find the covariance matrix $\mathbf{K}_{\mathbf{XX}}$.

[For parts (b) through (d), express your answers in terms of the matrix $\mathbf{A}$].

8.4 Let a collection of sequences $x(n, \theta_k)$ be given in terms of a deterministic parameter θ_k as

$$\left\{\cos(\frac{2\pi n}{5} + \theta_k)\right\}_{k=0}^{N-1}.$$

Now define a random variable Θ taking on values from the same parameter set $\{\theta_k\}$. Let the PMF of Θ be given as

$$P_\Theta(\theta_k) = \frac{1}{N} \quad \text{for } k = 0, ..., N-1.$$

Now set $X[n] \triangleq \cos(\frac{2\pi n}{5} + \Theta)$.

(a) Is $X[n]$ a random sequence? If so, describe both the mapping $X(n, \zeta)$ and its probability space (Ω, F, P). If not, explain fully.
(b) Let $\theta_k = \frac{2\pi k}{N}$ for $k = 0, ..., N-1$, and find $E\{X[n]\}$.†
(c) For the same θ_k as in part (b), find $E\{X[n]X[m]\}$. Take $N > 2$ here.

***8.5** Often one is given a problem statement starting as follows: "Let X be a real-valued random variable with pdf $f_X(x)\ldots$." Since an RV is a mapping from a sample space Ω with field of events $\mathscr{F}$ and a probability measure P, evidently the existence of an underlying probability space $(\Omega, \mathscr{F}, P)$ is assumed by such a problem statement. Show that a suitable underlying probability space $(\Omega, \mathscr{F}, P)$ can always be created, thus legitimizing problem statements such as the one above.

8.6 Let T be a continuous random variable denoting the time at which the first photon is emitted from a light source; T is measured from the instant the source is energized. Assume that the probability density function for T is $f_T(t) = \lambda e^{-\lambda t} u(t)$ with $\lambda > 0$.

---

†Note: $\cos(A + B) = \cos A \cos B - \sin A \sin B$ *and* $\cos A \cos B = \frac{1}{2}\{\cos(A + B) + \cos(A - B)\}$.

(a) What is the probability that at least one photon is emitted prior to time t_2 if it is known that none was emitted prior to time t_1, where $t_1 < t_2$?
(b) What is the probability that at least one photon is emitted prior to time t_2 if three independent sources of this type are energized simultaneously?

8.7 Let $Z_1, Z_2, \ldots$ be independent and identically distributed random variables with $P(Z_n = 1) = p$ and $P(Z_n = -1) = 1 - p, \forall n$

Let $X_n = \sum_{i=1}^{n} Z_i, n = 1, 2, \ldots$ and $X_0 = 0$. $\{X_n\}$ is called a simple random walk in one dimension

(a) Compute the first order pmf of X_n.
(b) Find $P(X_n = -2)$ after 4 steps.

8.8 Let X and Y be i.i.d. random variables with the exponential probability density functions

$$f_X(w) = f_Y(w) = \lambda e^{-\lambda w} u(w).$$

(a) Determine the probability density function for the ratio

$$0 \leq R \triangleq \frac{X}{X+Y} \leq 1, \quad \text{that is, } f_R(r), \qquad 0 < r \leq 1.$$

(b) Let A be the event $X < 1/Y$. Determine the conditional pdf of X given that A occurs and that $Y = y$; that is, determine

$$f_X(x|A, Y = y).$$

(c) Using the definitions of (b), what is the minimum mean-square error estimate of X given that the event A occurs and that $Y = y$?

8.9 Use the Schwarz inequality for complex random variables to prove that

$$|R_X[m]| \leq R_X[0], \qquad \text{for all integers } m$$

for any complex-valued WSS random sequence $X[m]$.

8.10 Let $\mathbf{X} = (X_1, X_2, \ldots, X_{10})^T$ be a random vector whose components satisfy the equations,

$$X_i = \frac{2}{5}(X_{i-1} + X_{i+1}) + W_i \quad \text{for } 2 \leq i \leq 9,$$

where the W_i are independent and Laplacian distributed with mean zero and variance σ^2 for $i = 1$ to 10, and $X_1 = \frac{1}{2}X_2 + \frac{5}{4}W_1$ and $X_{10} = \frac{1}{2}X_9 + \frac{5}{4}W_{10}$.

(a) Find the mean vector $\mu_{\mathbf{X}}$.
(b) Find the covariance matrix $\mathbf{K_{XX}}$.
(c) Write an expression for the multidimensional pdf of the random vector $\mathbf{X}$.

[*Hint*:

$$\text{Matrix identity:} \qquad \text{if } \mathbf{A} \triangleq \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^N \\ \rho & 1 & \rho & \rho^2 & \dots \\ \rho^2 & \rho & 1 & \dots & \dots \\ \dots & \rho^2 & \dots & \dots & \rho \\ \rho^N & \dots & \dots & \rho & 1 \end{bmatrix},$$

then $\mathbf{A}^{-1}$ is given as

$$\beta^2 \mathbf{A}^{-1} = \begin{bmatrix} 1-\rho\alpha & -\alpha & 0 & \dots & 0 \\ -\alpha & 1 & -\alpha & 0 & \dots \\ 0 & -\alpha & 1 & \dots & 0 \\ \dots & 0 & \dots & \dots & -\alpha \\ 0 & \dots & 0 & -\alpha & 1-\rho\alpha \end{bmatrix}$$

with $\alpha \triangleq \frac{\rho}{1+\rho^2}$ and $\beta^2 \triangleq \frac{1-\rho^2}{1+\rho^2}$. The Laplacian pdf is given as

$$f_W(w) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\sqrt{2}\frac{|w|}{\sigma}\right), \qquad -\infty \le w \le +\infty.$$

8.11 Prove Corollary 8.1-1.

8.12 Let $\{X_i\}$ be a sequence of i.i.d. Normal random variables with zero-mean and unit variance. Let

$$S_k \triangleq X_1 + X_2 + \ldots + X_k \quad \text{for } k \ge 1.$$

Determine the joint probability density function for S_n and S_m, where $1 \le m < n$.

8.13 In Example 8.1-8 we saw that CDFs are continuous from the right. Are they continuous from the left also? Either prove or give a counterexample.

8.14 Let the probability space $(\Omega, \mathscr{F}, P)$ be given as follows:

$$\Omega = \{a, b, c\}, \text{that is, the outcome } \zeta = a \text{ or } b \text{ or } c,$$

$$\mathscr{F} = \text{all subsets of } \Omega,$$

$$P[\{\zeta\}] = 1/3 \text{ for each outcome } \zeta.$$

Let the random sequence $X[n]$ be defined as follows:

$$X[n, a] = 3\delta[n]$$

$$X[n, b] = u[n-1]$$

$$X[n, c] = \cos \pi n/2.$$

(a) Find the mean function $\mu_X[n]$.
(b) Find the correlation function $R_{XX}[m, n]$.
(c) Are $X[1]$ and $X[0]$ independent? Why?

8.15 Let X_n be an independently and identically distributed sequence of Gaussian random variables with zero mean and variance σ^2. Let Y_n be the average of the two consecutive values of X_n:

$$Y_n = \frac{X_n + X_{n-1}}{2}$$

Determine whether $\{Y_n\}$ is a wide-sense stationary process.

8.16 Consider a random sequence $X[n]$ as the input to a linear filter with impulse response

$$h[n] = \begin{cases} 1/2, & n = 0 \\ 1/2, & n = 1 \\ 0, & \text{else.} \end{cases}$$

We denote the output random sequence $Y[n]$, that is, for each outcome ζ,

$$Y[n, \zeta] = \sum_{k=-\infty}^{k=+\infty} h[k]X[n-k, \zeta].$$

Assume the filter runs for all time, $-\infty < n < +\infty$. We are given the mean function of the input $\mu_X[n]$ and correlation function of the input $R_{XX}[n_1, n_2]$. Express your answers in terms of these assumed known functions.

(a) Find the mean function of the output $\mu_Y[n]$.
(b) Find the autocorrelation function of the output $R_{YY}[n_1, n_2]$.
(c) Write the autocovariance function of the output $K_{YY}[n_1, n_2]$ in terms of your answers to parts (a) and (b).
(d) Now assume that the input $X[n]$ is a Gaussian random sequence, and write the corresponding joint pdf of the output $f_Y(y_1, y_2; n_1, n_2)$ at two arbitrary times $n_1 \neq n_2$ in terms of $\mu_Y[n]$ and $K_{YY}[n_1, n_2]$.

8.17 Let I_n be a sequence of independent Bernoulli random variables. I_n is then an independent and identically distributed random sequence taking on values from the set $\{0, 1\}$.

(a) Let $D_n = 2I_n - 1$. Then

$$D_n = \begin{cases} 1, & I_n = 1 \\ -1, & I_n = 0 \end{cases}$$

Determine the mean and variance of D_n.

(b) Let $S_n = \sum_{k=1}^{n} I_k$. Obtain the first order pmf of S_n and its mean and variance.

8.18 Let $T[n]$ denote the random arrival sequence studied in class,

$$T[n] = \sum_{k=1}^{n} \tau[k],$$

where the $\tau[k]$ are an *independent random sequence* of interarrival times, distributed as exponential with parameter $\lambda > 0$.

(a) Find the CF of this random sequence, that is,

$$\Phi_T(\omega; n) = E[e^{j\omega T[n]}].$$

(b) Use this CF to find the mean function $\mu_T[n]$.

8.19 Let the random sequence $T[n]$ be defined on $n \geq 1$ and for each n, have an Erlang pdf:

$$f_T(t; n) = \frac{(\lambda t)^{n-1}}{(n-1)!}\lambda e^{-\lambda t}u(t), \quad \lambda > 0.$$

Define the new random sequence $\tau[n] \triangleq T[n] - T[n-1]$ for $n \geq 2$, and set $\tau[1] \triangleq T[1]$. Can we conclude that $\tau[n]$ is exponential with the same parameter λ? If not, what additional information on the random sequence $T[n]$ is needed? Justify your answer.

8.20 This problem considers a random sequence model for a *charge coupled device* (CCD) array with very "leaky" cells. We start by defining the width-3 pulse function:

$$h[n] = \begin{cases} 1/4 & n = -1 \\ 1/2 & n = 0 \\ 1/4 & n = +1 \\ 0 & \text{else,} \end{cases}$$

and as illustrated in Figure P8.20, which we will use to account for 25 percent of the charge in a cell that leaks out to its right neighbor and 25 percent that leaks to its left neighbor. We assume that the one-dimensional CCD array is infinitely long and represents the array contents by the random sequence X:

$$X[n, \zeta] = \sum_{i=-\infty}^{i=+\infty} A(\zeta_i)h[n-i],$$

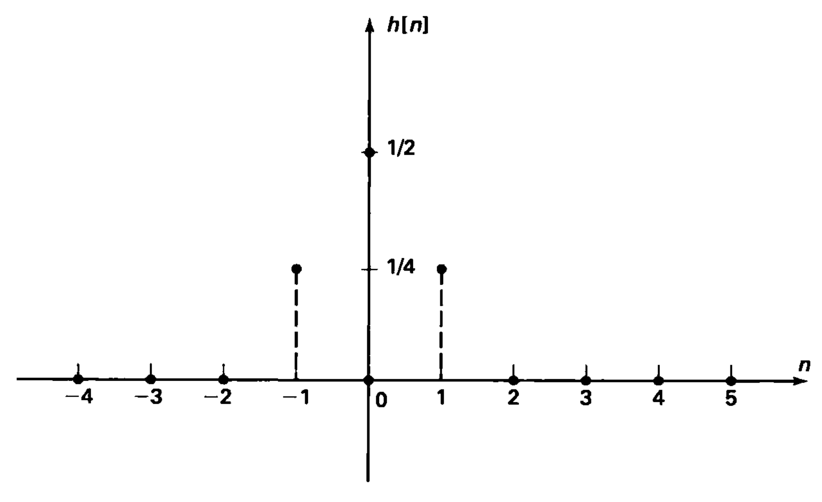


Figure P8.20 Pulse function of leaky cell.

where ζ_i is the ith component of ζ, the infinite dimensional outcome of the experiment. The random variables $A(\zeta_i)$ are jointly independent and Gaussian distributed with mean λ and variance λ.

(a) Find the mean function $\mu_X[n]$.
(b) Find the first-order pdf $f_X(x;n)$.
(c) Find the joint pdf $f_X(x_1, x_2; n, n+1)$.

8.21 We are given a random sequence $X[n]$ for $n \geq 0$ with conditional pdf's

$$f_X(x_n|x_{n-1}) = \alpha \exp[-\alpha(x_n - x_{n-1})]\, u(x_n - x_{n-1}) \quad \text{for } n \geq 1,$$

with $u(x)$ the unit-step function and initial pdf $f_X(x_0) = \delta(x_0)$. Take $\alpha > 0$.

(a) Find the first-order pdf $f_X(x_n)$ for $n = 2$.
(b) Find the first-order pdf $f_X(x_n)$ for arbitrary $n > 1$ using mathematical induction.

8.22 Let $x[n]$ be a deterministic input to the LSI discrete-time system H shown in Figure P8.23.

(a) Use linearity and shift-invariance properties to show that

$$y[n] = x[n] * h[n] \triangleq \sum_{k=-\infty}^{+\infty} x[k]h[n-k] = h[n] * x[n].$$

(b) Define the Fourier transform of a sequence $a[n]$ as

$$A(\omega) \triangleq \sum_{n=-\infty}^{\infty} a[n]e^{-j\omega n}, \qquad -\pi \leq \omega \leq +\pi,$$

and show that the inverse Fourier transform is

$$a[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} A(\omega)e^{+j\omega n}d\omega, \qquad -\infty < n < +\infty.$$

(c) Using the results in (a) and (b), show that

$$Y(\omega) = H(\omega)X(\omega), \qquad -\pi \leq \omega \leq +\pi,$$

for an LSI discrete-time system.

8.23 Consider the difference equation

$$y[n] + \alpha y[n-1] = x[n], \qquad -\infty < n < +\infty,$$

where $-1 < \alpha < +1$.

(a) Let the input be $x[n] = \beta^n u[n]$ for $-1 < \beta < +1$. Find the solution for $y[n]$ assuming causality applies, that is, $y[n] = 0$ for $n < 0$.

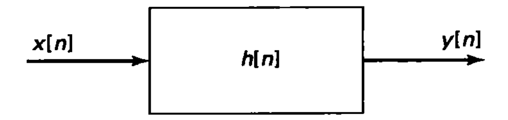


Figure P8.23 LSI system with impulse response $h[n]$.

(b) Let the input be $x[n] = \beta^{-n}u[-n]$ for $-1 < \beta < +1$. Find the solution for $y[n]$ assuming anticausality applies,† that is, $y[n] = 0$ for $n > 0$.

8.24 Let $X[n]$ be a WSS random sequence with mean zero and covariance function

$$K_{XX}[m] = \sigma^2 \rho^{|m|} \quad \text{for all} \quad -\infty < m < +\infty,$$

where ρ is a real constant. Consider difference equations of the form

$$Y[n] = X[n] - \alpha X[n-1] \quad \text{with} \quad -\infty < n < +\infty.$$

(a) Write the covariance function of $Y[n]$ in terms of the parameters σ^2, ρ, and α.
(b) Find a value of α such that $Y[n]$ is a WSS white noise sequence.
(c) What is the average power of this white noise?

8.25 Let $W[n]$ be an independent random sequence with mean 0 and variance σ_W^2 defined for $-\infty < n < +\infty$. For appropriately chosen ρ, let the stationary random sequence $X[n]$ satisfy the causal LCCDE

$$X[n] = \rho X[n-1] + W[n], \qquad -\infty < n < +\infty.$$

(a) Show that $X[n-1]$ and $W[n]$ are independent at time n.
(b) Derive the characteristic function equation

$$\Phi_X(\omega) = \Phi_X(\rho\omega)\Phi_W(\omega).$$

(c) Find the continuous solution to this functional equation for the unknown function Φ_X when $W[n]$ is assumed to be Gaussian. [*Note:* $\Phi_X(0) = 1$.]
(d) What is σ_X^2?

†This part requires more detailed knowledge of the z-transform. (cf. Appendix A.)

8.26 Consider the LSI system shown in Figure P8.26, whose *deterministic input* $x[n]$ is contaminated by noise (a random sequence) $W[n]$. We wish to determine the properties of the output random sequence $Y[n]$. The noise $W[n]$ has mean $\mu_W[n] = 2$ and autocorrelation $E\{W[m]W[n]\} = \sigma_W^2\delta[m-n] + 4$. The impulse response is $h[n] = \rho^n u[n]$ with $|\rho| < 1$. The deterministic input $x[n]$ is given as $x[n] = 3$ for all n.

(a) Find the output mean $\mu_Y[n]$.
(b) Find the output power $E\{Y^2[n]\}$.
(c) Find the output covariance $K_{YY}[m,n]$.

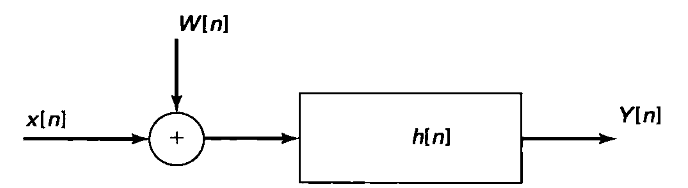


Figure P8.26 LSI system with deterministic-plus-noise input.

8.27 Show that the random sequence $X[n]$ generated in Example 8.1-15 is not an independent random sequence.

8.28 The impulse response of a discrete linear time-invariant system is given by $h[n] = a^n u[n]$ where $|a| < 1$, and $u[n]$ is the unit step sequence defined by

$$u[n] = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases}$$

If the input sequence $X[n]$ is a discrete-time white noise with power spectral density $\frac{N_0}{2}$, find the power spectral density of the output $Y[n]$.

8.29 Let X_n consist or two interleaved sequences of independent random variables. For n even, X_n assumes the values ± 1 with probability $\frac{1}{2}$; for n odd, X_n assumes the values $\frac{1}{3}$ and -3 with probabilities $\frac{9}{10}$ and $\frac{1}{10}$ respectively. verify whether

(a) $\{X_n\}$ is WSS.
(b) $\{X_n\}$ is stationary.

8.30 Consider a WSS random sequence $X[n]$ with mean $\mu_X[n] = \mu$, a constant, and correlation function $R_{XX}[m] = p^2\delta[m]$ with $p^2 > 0$. In such a case μ must be zero, as you will show in this problem. Note that the covariance function here is $K_{XX}[m] = p^2\delta[m] - \mu^2$.

(a) Take $m = 0$ and conclude that $p^2 \geq \mu^2$.
(b) Take a vector $\mathbf{X}$ of length N out of the random sequence $X[n]$. Show that the corresponding covariance matrix $\mathbf{K_{XX}}$ will be positive semidefinite only if $\mu^2 \leq \sigma^2/(N-1)$, where $\sigma^2 \triangleq p^2 - \mu^2$. (*Hint:* Take coefficient vector $\mathbf{a} = \mathbf{1}$, i.e., all 1's.)
(c) Let $N \to \infty$ and conclude that μ must be zero for the stationary white noise sequence $X[n]$.

8.31 A discrete-time system is given by

$$Y[n] = aY[n-1] + X[n] \text{ where } |a| < 1.$$

The input $X[n]$ is discrete-time white noise with average σ^2. The impulse response $h[n]$ of the system defined by $h[n] = ah[n-1] + \delta[n]$. Find the spectral density and average power of the output $Y[n]$.

8.32 Let the WSS random sequence X have correlation function

$$R_{XX}[m] = 10e^{-\lambda_1|m|} + 5e^{-\lambda_2|m|}$$

with $\lambda_1 > 0$ and $\lambda_1 > 0$. Find the corresponding psd $S_{XX}(\omega)$ for $|\omega| \leq \pi$.

8.33 The psd of a certain random sequence is given as $S_{XX}(\omega) = 1/[(1+\alpha^2) - 2\alpha\cos\omega]^2$ for $-\pi \leq \omega \leq +\pi$, where $|\alpha| < 1$. Find the random sequence's correlation function $R_X[m]$.

8.34 Let the input to system $H(\omega)$ be $W[n]$, a white noise random sequence with $\mu_W[n] = 0$ and $K_{WW}[m] = \delta[m]$. Let $X[n]$ denote the corresponding output random sequence. Find $K_{XW}[m]$ and $S_{XW}(\omega)$.

8.35 Consider the system shown in Figure P8.35. Let $X[n]$ and $V[n]$ be WSS and mutually uncorrelated with zero mean and psd's $S_{XX}(\omega)$ and $S_{VV}(\omega)$, respectively.

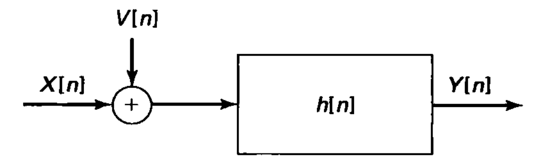


Figure P8.35 LSI system with random signal-plus-noise input.

(a) Find the psd of the output $S_{YY}(\omega)$.
(b) Find the cross-power spectral density between the input X and the output Y, that is, find $S_{XY}(\omega)$.

8.36 Consider the discrete-time system with input random sequence $X[n]$ and output $Y[n]$ given as

$$Y[n] = \frac{1}{5}\sum_{k=-2}^{+2} X[n-k].$$

Assume that the input sequence $X[n]$ is WSS with psd $S_{XX}(\omega) = 2$.

(a) Find the psd of the output random sequence $S_{YY}(\omega)$.
(b) Find the output correlation function $R_{YY}[m]$.

8.37 Let the stationary random sequence $Y[n] = X[n]+U[n]$ with power spectral density (psd) $S_Y(\omega)$ be our model of signal X plus noise U for a certain discrete-time channel. Assume that X and U are orthogonal and also assume that we have $S_Y(\omega) > 0$ for all $|\omega| \leq \pi$. As a first step in processing Y to find an estimate for X, let Y be input to a discrete-time filter $G(\omega)$ defined as $G(\omega) = 1/\sqrt{S_Y(\omega)}$ to produce the stationary output sequence $W[n]$ as shown in Figure P8.37a.

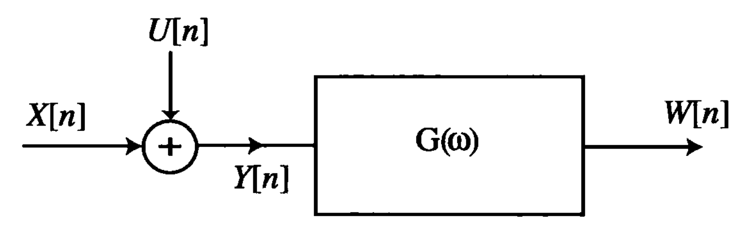


Figure P8.37a

(a) Find the psd of $W[n]$, that is, $S_W(\omega)$, and also the cross-power spectral density between original input and output $S_{XW}(\omega)$, in terms of S_X, S_Y, and S_U.
(b) Next filter $W[n]$ with an FIR impulse response $h[n]$, $n = 0, ..., N-1$, to give output $\widehat{X}[n]$, an estimate of the original noise-free signal $X[n]$ as shown in Figure P8.37b. In line with the Hilbert space theory of random variables, we

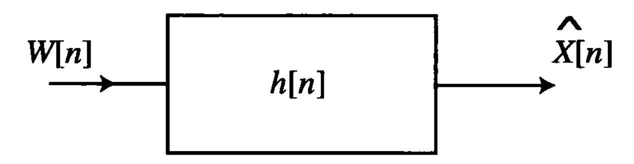


Figure P8.37b

decide to choose the filter coefficients $h[n]$ so that the estimate error $\widehat{X}[n] - X[n]$ will be orthogonal to all those $W[n]$ actually used in making the estimate at time n. Write down the resulting equations for the N filter coefficients $h[0], h[1], ..., h[N-1]$. Your answer should be in terms of the cross-correlation function $R_{XW}[m]$.
(c) Let N go to infinity, and write the frequency response of $h[n]$, that is, $H(\omega)$, in terms of the discrete-time power spectral densities $S_{XX}(\omega)$ and $S_{YY}(\omega)$.

8.38 Higher than second-order moments have proved useful in certain advanced applications. Here we consider a third-order correlation function of a stationary random sequence

$$R_X[m_1, m_2] \triangleq E\{X[n+m_1]X[n+m_2]X^*[n]\}$$

defined for the random sequence $X[n]$, $-\infty < n < +\infty$.

(a) Let $Y[n]$ be the output from an LSI system with impulse response $h[n]$, due to the input random sequence $X[n]$. Determine a convolution-like equation expressing the third-order correlation function of the output $R_Y[m_1, m_2]$ in terms of the third-order correlation function of the input $R_X[m_1, m_2]$ and the system impulse response $h[n]$.

Define the bi-spectral density of X as the two-dimensional Fourier transform

$$S_X(\omega_1, \omega_2) = \sum_{m_1}\sum_{m_2} R_X[m_1, m_2] \exp -j(\omega_1 m_1 + \omega_2 m_2).$$

(b) For the system of part (a), find an expression for the bi-spectral density of the output $S_Y(\omega_1, \omega_2)$ in terms of the system frequency response $H(\cdot)$ and the bi-spectral density of the input $S_X(\omega_1, \omega_2)$.

8.39 Let $X[n]$ be a Markov chain on $n \geq 0$ taking values 1 and 2 with *one-step transition probabilities*,

$$\mathbf{P}_{ij} \triangleq P\{X[n] = j | X[n-1] = i\}, \qquad 1 \leq i, j \leq 2,$$

given in matrix form as

$$\mathbf{P} = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = (p_{i,j}).$$

We describe the state probabilities at time n by the vector

$$\mathbf{p}[n] \triangleq [P\{X[n] = 1\}, P\{X[n] = 2\}].$$

(a) Show that $\mathbf{p}[n] = \mathbf{p}[0]\mathbf{P}^n$.
(b) Draw a two-state transition diagram and label the branches with the one-step transition probabilities p_{ij}. Don't forget the p_{ii} or *self-transitions*. (See Figure 8.5-1 for state-transition diagram of a Markov chain.)
(c) Given that $X[0] = 1$, find the probability that the first transition to state 2 occurs at time n.

8.40 Consider using a first-order Markov sequence to model a random sequence $X[n]$ as

$$X[n] = rX[n-1] + Z[n],$$

where $Z[n]$ is white noise of variance σ_Z^2. Thus, we can look at $X[n]$ as the output of passing $Z[n]$ through a linear system. Take $|r| < 1$ and assume the system has been running for a long time, that is, $-\infty < n < +\infty$.

(a) Find the psd of $X[n]$, that is, $S_{XX}(\omega)$.
(b) Find the correlation function $R_{XX}[m]$.

8.41 We defined a Markov random sequence $X[n]$ in this chapter as being specified by its first-order pdf $f_X(x; n)$ and its one-step conditional pdf

$$f_X(x_n | x_{n-1}; n, n-1) = f_X(x_n | x_{n-1}) \quad \text{for short.}$$

(a) Find the two-step pdf for a Markov random sequence $f_X(x_n|x_{n-2})$ in terms of the above functions. Here, take $n \geq 2$ for a random sequence starting at $n = 0$.

(b) Find the N-step pdf $f_X(x_n|x_{n-N})$ for arbitrary positive integer N, where we only need consider $n \geq N$.

8.42 Consider a generalized random walk sequence $X[n]$ running on $\{n \geq 0\}$ and defined as follows:

$$X[0] \triangleq 0,$$

$$X[n] \triangleq \sum_{k=1}^{n} W[k], \quad n \geq 0,$$

where $W[n]$ is an *independent random sequence*, stationary, and taking values below with the indicated probabilities,

$$W[n] \triangleq \begin{cases} +s_1, p = 1/2, \\ -s_2, p = 1/2. \end{cases}$$

We see the difference is that the positive and negative step sizes are not the same $s_1 \neq s_2, s_1 > 0$ and $s_2 > 0$.

(a) Find the mean function $\mu_X[n] \triangleq E\{X[n]\}$.

(b) Find the autocorrelation function $R_X[n_1, n_2] \triangleq E\{X[n_1]X[n_2]\}$.

8.43 Consider a Markov random sequence $X[n]$ running on $1 \leq n \leq 100$. It is statistically described by its first-order pdf $f_X(x;1)$ and its one-step transition (conditional) pdf $f_X(x_n|x_{n-1}; n, n-1)$. By the Markov definition, we have (suppressing the time variables) that

$$f_X(x_n|x_{n-1}) = f_X(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) \quad \text{for } 2 \leq n \leq 100.$$

Show that a Markov random sequence is also Markov *in the reverse order*, that is,

$$f_X(x_n|x_{n+1}) = f_X(x_n|x_{n+1}, x_{n+2}, \ldots, x_{100}) \quad \text{for } 1 \leq n \leq 99,$$

and so one can alternatively statistically describe a Markov random sequence by the one-step backward pdf $f_X(x_{n-1}|x_n; n-1, n)$ and first-order pdf $f_X(x; 100)$.

8.44 Suppose that the probability of a sunny day (state 0) following a rainy day (state 1) is $\frac{1}{3}$, and that the probability of a rainy day following a sunny day is $\frac{1}{2}$. Write the 2-state transition probability matrix. Given that May 1 is a suuny day, determine the probability that May 3 is a sunny day and May 5 is a suuny day.

8.45 Consider the Markov random sequence $X[n]$ generated by the difference equation, for $n \geq 1$,

$$X[n] = \alpha X[n-1] + \beta W[n],$$

where the input $W[n]$ is an *independent random sequence* with zero mean and variance σ_W^2, the inital value $X[0] = 0$, and the parameters α and β are known constants.

(a) Show that the subsequence $Y[n] \triangleq X[2n]$ is Markov also.
(b) Find the variance function $\sigma_Y^2[n] \triangleq E[|Y[n] - \mu_Y[n]|^2]$ for $n > 0$.

***8.46** Write a MATLAB function called `triplemarkov` that will compute and plot the autocorrelation functions for the asymmetric, two-state Markov model in Example 8.1-16 for any three sets of parameters $\{p_{00}, p_{11}\}$. Denote the maximum lag interval as N. Run your routine for $\{0.2, 0.8\}, \{0.2, 0.5\}$, and $\{0.2, 0.2\}$. Repeat for $\{0.8, 0.2\}$, $\{0.8, 0.5\}$, and $\{0.8, 0.8\}$. Describe what you observe.

8.47 Consider the probability space $(\Omega, \mathscr{F}, P)$ with $\Omega = [0, 1]$, $\mathscr{F}$ defined to be the Borel sets of Ω, and $P[(0, \zeta] = \zeta$ for $0 < \zeta \leq 1$.

(a) Show that $P[\{0\}] = 0$ by using the axioms of probability.
(b) Determine in what *senses* the following random sequences converge:
   (i) $X[n, \zeta] = e^{-n\zeta}, n \geq 0$
   (ii) $X[n, \zeta] = \sin\left(\zeta + \frac{1}{n}\right), n \geq 1$
   (iii) $X[n, \zeta] = \cos^n(\zeta), n \geq 0$.
(c) If the preceding sequences converge, what are the limits?

8.48 The members of the sequence of jointly independent random variables $X[n]$ have pdf's of the form

$$f_X(x; n) = \left(1 - \frac{1}{n}\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(x - \frac{n-1}{n}\sigma\right)^2\right] + \frac{1}{n}\sigma \exp(-\sigma x) u(x).$$

Determine whether or not the random sequence $X[n]$ converges in

(i) the mean-square sense,
(ii) probability,
(iii) distribution.

8.49 The members of the random sequence $X[n]$ have joint pdf's of the form

$$f_X(\alpha, \beta; m, n) = \frac{mn}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1 - \rho^2)}[m^2\alpha^2 - 2\rho mn\alpha\beta + n^2\beta^2]\right)$$

for $m \geq 1$ and $n \geq 1$ where $-1 < \rho < +1$.

(a) Show that $X[n]$ converges in the mean-square sense as $n \to \infty$ for all $-1 < \rho < +1$.
(b) Specify the CDF of the mean-square limit $X \triangleq \lim_{n\to\infty} X[n]$.

8.50 State conditions under which the mean-square limit of a sequence of Gaussian random variables is also Gaussian.

8.51 Let $X[n]$ be a real-valued random sequence on $n \geq 0$, made up from stationary and *independent increments*, that is, $X[n] - X[n-1] = W[n]$, "the increment" with $W[n]$ being a stationary and independent random sequence. The random sequence

always starts with $X[0] = 0$. We also know that at time $n = 1$, $E\{X[1]\} = \eta$ and $\text{Var}\{X[1]\} = \sigma^2$.

(a) Find $\mu_X[n]$ and $\sigma_X^2[n]$, the mean and variance functions of the random sequence X at time n for any time $n > 1$.

(b) Prove that $X[n]/n$ converges in probability to η as the time n approaches infinity.

8.52 This problem demonstrates that p-convergence implies convergence in distribution even when the limiting pdf does not exist.

(a) For any real number x and any positive ε, show that

$$P[X \le x - \varepsilon] \le P[X[n] \le x] + P[|X[n] - X| \ge \varepsilon].$$

(b) Similarly show that

$$P[X > x + \varepsilon] \le P[X[n] > x] + P\left[|X[n] - X| \ge \varepsilon\right].$$

For part (c), assume the random sequence $X[n]$ converges to the random variable X in probability.

(c) Let $n \to \infty$ and conclude that

$$\lim_{n \to \infty} F_X(x; n) = F_X(x)$$

at points of continuity of F_X.

8.53 Let $X[n]$ be a second-order random sequence. Let $h[n]$ be the impulse response of an LSI system. We wish to define the output of the system $Y[n]$ as a mean-square limit.

(a) Show that we can define the mean-square limit

$$Y[n] \triangleq \sum_{k=-\infty}^{+\infty} h[k]X[n-k], \qquad -\infty < n < +\infty, \ \text{(m.s.)}$$

if

$$\sum_k \sum_l h[k]h^*[l]R_{XX}[n-k, n-l] < \infty \text{ for all } n.$$

(*Hint:* Set $Y_N[n] \triangleq \sum_{k=-N}^{+N} h[k]X[n-k]$ and show that m.s. limit of $Y_N[n]$ exists by using the Cauchy convergence criteria.)

(b) Find a simpler condition for the case when $X[n]$ is a wide-sense stationary random sequence.

(c) Find the necessary condition on $h[n]$ when $X[n]$ is (stationary) white noise.

8.54 If $X[n]$ is a Martingale sequence on $n \ge 0$, show that

$$E\{X[n+m]|X[m], \ldots, X[0]\} = X[m] \qquad \text{for all } n \ge 0.$$

8.55 Let $Y[n]$ be a random sequence and X a random variable and consider the conditional expectation

$$E\{X|Y[0],\ldots,Y[n]\} \triangleq G[n].$$

Show that the random sequence $G[n]$ is a Martingale.

***8.56** We can enlarge the concept of Martingale sequence somewhat as follows. Let $G[n] \triangleq g(X[0],\ldots,X[n])$ for each $n \geq 0$ for measurable functions g. We say G is a *Martingale* with respect to X if $E\{G[n]|X[0],\ldots,X[n-1]\} = G[n-1]$.

(a) Show that Theorem 8.8-3 holds for G a Martingale with respect to X. Specifically, substitute G for X in the statement of the theorem. Then make necessary changes to the proof.

(b) Show that the Martingale convergence Theorem 8.8-4 holds for G a Martingale with respect to X.

8.57 Consider the hypothesis-testing problem involving $(n+1)$ observations $X[0],\ldots,X[n]$ of the random sequence X. Define the likelihood ratio

$$L_X[n] \triangleq \frac{f_X(X[0],\ldots,X[n]|H_1)}{f_X(X[0],\ldots,X[n]|H_0)}, \qquad n \geq 0,$$

corresponding to two hypotheses H_1 and H_0. Show that $L_X[n]$ is a Martingale with respect to X under hypothesis H_0.

8.58 In the discussion of interpolation in Example 8.4-7, work out the algebra needed to arrive at the psd of the up-sampled random sequence $X_e[n]$.

8.59 The up-sampled sequence $X_e[n]$ in the interpolation process is clearly not WSS, even if $X[n]$ is WSS. Create an up-sampled random sequence that is WSS by randomizing the start-time of the sequence $X[n]$. That is, define a binary random variable Θ with $P[\Theta = 0] = P[\Theta = 1] = 0.5$. Define the start-time randomized sequence by $X_r[n] \triangleq X[n+\Theta]$. Then the resulting up-sampled sequence is $X_{er}[n] = X\left[\frac{n+\Theta}{2}\right]$. Show that $R_{X_rX_r}[k] = R_{XX}[k]$ and $R_{X_{er}X_{er}}[m, m+k] = R_{X_{er}X_{er}}[k] = 0.5R_{XX}[k/2]$ for k even, and zero for k odd.

REFERENCES

8-1. R. B. Ash, *Real Analysis and Probability.* New York: Academic Press, 1972, pp. 1–53 and 115.

8-2. A. Kolmogorov, *Foundations of the Theory of Probability.* New York: Chelsea, 1950.

8-3. T. M. Apostol, *Mathematical Analysis.* Reading, MA: Addison-Wesley, 1957, pp. 192–202.

8-4. Y. Viniotis, *Probability and Random Processes.* Boston, MA: WCB/Mcgraw-Hill, 1998, p. 103.

8-5. A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing,* 2nd Edition. Upper Saddle River, NJ: Prentice Hall, 1999, Chapters 2–3.

8-6. H. Stark and Y. Yang, *Vector Space Projections.* New York: Wiley, 1998.

8-7. S. S. Soliman and M. D. Srinath, *Continuous and Discrete Signals and Systems.* Upper Saddle River, NJ: Prentice Hall, 1998.

8-8. W. Rudin, *Principles of Mathematical Analysis.* New York: McGraw-Hill, 1964, pp. 45–48.

8-9. B. V. Gnedenko, *The Theory of Probability* (translated by B. D. Seckler). New York: Chelsea, 1963, p. 237.

8-10. A. Leon-Garcia, *Probability, Statistics, and Random processes for Electrical Engineering,* 3rd edition. Upper Saddle River, NJ: Prentice Hall, 2008.

8-11. G. Grimmett and D. Strizaker, *Probability and Random Processes*, 3rd edition. Oxford, England: Oxford University Press, 2001.

9 Random Processes

In the last chapter, we learned how to generalize the concept of random variable to that of random sequence. We did this by associating a sample sequence with each outcome $\zeta \in \Omega$, thereby generating a family of sequences collectively called a random sequence. These sequences were indexed by a discrete (integer) parameter n is some index set Z. In this chapter we generalize further by considering random functions of a *continuous* parameter. We consider this continuous parameter time, but it could equally well be position, or angle, or some other continuous parameter. The collection of all these continuous time functions is called a random process. Random processes will be perhaps the most useful objects we study because they can be used to model physical processes directly without any intervening need to sample the data. Even when of necessity one is dealing with sampled data, the concept of random process will give us the ability to reference the properties of the sample sequence to those of the limiting continuous process so as to be able to judge the adequacy of the sampling rate.

Random processes find a wide variety of applications. Perhaps the most common use is as a model for noise in physical systems, modeling of the noise being the necessary first step in deciding on the best way to mitigate its negative effects. A second class of applications concerns the modeling of random phenomena that are not noise but are nevertheless unknown to the system designer. An example would be a multimedia signal (audio, image, or video) on a communications link. The signal is not noise, but it is unknown from the viewpoint of a distant receiver and can take on many (an enormous number of) values. Thus, we model such signals as random processes, when some statistical description of the source is available. Situations such as this arise in other contexts also, such as control systems,

pattern recognition, etc. Indeed from an information theory viewpoint, any waveform that communicates information must have at least some degree of randomness in it.

We start with a definition of random process and study some of the new difficulties to be encountered with continuous time. Then we look at the moment functions for random processes and generalize the correlation and covariance functions from Chapter 8 to this continuous parameter case. We also look at some basic random processes of practical importance. We then begin a study of linear systems and random processes. Indeed, this topic is central to our study of random processes and is widely used in applications. Then we present some classifications of random processes based on general statistical properties. Finally, we introduce stationary and wide-sense stationary random processes and their analysis for linear systems.

9.1 BASIC DEFINITIONS

It is most important to fully understand the basic concept of the random process and its associated moment functions. The situation is analogous to the discrete-time case treated in Chapter 8. The main new difficulty is that the time axis has now become uncountable. We start with the basic definition.

Definition 9.1-1 Let $(\Omega, \mathscr{F}, P)$ be a probability space. Then define a mapping X from the sample space Ω to a space of continuous time functions. The elements in this space will be called *sample functions*. This mapping is called a *random process* if at each fixed time the mapping is a random variable, that is, $X(t,\zeta) \in \mathscr{F}$ † for each fixed t on the real line $-\infty < t < +\infty$. ■

Thus we have a multidimensional function $X(t,\zeta)$, which for each fixed outcome ζ is an ordinary time function and for each fixed t is a random variable. This is shown diagrammatically in Figure 9.1-1 for the special case where the sample space Ω is the continuous interval $[0, 10]$. We see a family of random variables indexed by t when we look along the time axis, and we see a family of time functions indexed by ζ when we look along the outcome "axis."

We have the following elementary examples of random processes:

Example 9.1-1

(*simple process*) $X(t,\zeta) = X(\zeta)f(t)$, where X is a random variable and f is a deterministic function of the parameter t. We also write $X(t) = Xf(t)$.

Example 9.1-2

(*random sinewave*) $X(t,\zeta) = A(\zeta)\sin(\omega_0 t + \Theta(\zeta))$, where A and Θ are random variables. We also write $X(t) = A\sin(\omega_0 t + \Theta)$, suppressing the outcome ζ.

More typical examples of random processes can be constructed from random sequences.

† $X \in \mathscr{F}$ is shorthand for $\{\zeta\colon X(\zeta) \leq x\} \subset \mathscr{F}$ for all x. This condition permits us to measure the probability of events of this kind and hence define CDFs.

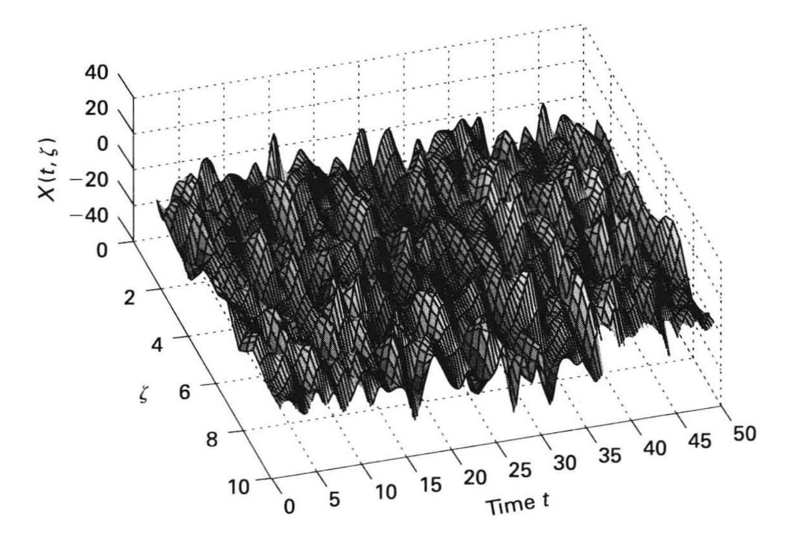


Figure 9.1-1 A random process for a continuous sample space $\Omega = [0,10]$.

Example 9.1-3

$X(t) = \sum_n X[n]p_n(t - T[n])$, where $X[n]$ and $T[n]$ are random sequences and the functions $p_n(t)$ are deterministic waveforms that can take on various shapes. For example, the $p_n(t)$ might be ideal unit-step functions that could provide a model for a so-called *jump process.* In this interpretation the $T[n]$ would be the times of the *arrivals* and the $X[n]$ would be the *amplitudes* of the jumps. Then $X(t)$ would indicate the total amplitude up to time t. If all the $X[n]$'s were 1, we would have a *counting process* in that $X(t)$ would count the arrivals prior to time t.

If we sample the random process at n times t_1 through t_n, we get an n-dimensional random vector. If we know the probability distribution of this vector for all times t_1 through t_n and for all positive n, then clearly we know a lot about the random process. If we know all this information, we say that we have *statistically specified* (statistically determined) the random process in a fashion that is analogous to the corresponding case for random sequences.

Definition 9.1-2 A random process $X(t)$ is statistically specified by its complete set of nth-order CDFs (pdf's or PMFs) for all positive integers n, that is, $F_X(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n)$ for all $x_1, x_2, \ldots, x_n$ and for all $-\infty < t_1 < t_2 < \ldots < t_n < \infty$. ■

The term *statistical* comes from the fact that this is the limit of the information that could be obtained from accumulating relative frequencies of events determined by the random process $X(t)$ at all finite collections of time instants. Clearly, this is all we could hope to determine by measurements on a process that we wish to model. However, the question arises: Is this enough information to completely determine the random process? Unfortunately the general answer is no. We need to impose a continuity requirement on the sample functions $x(t)$. To see this the following simple example suffices.

Example 9.1-4

(*from Karlin* [9-1]) Let U be a uniform random variable on $[0, 1]$ and define the random processes $X(t)$ and $Y(t)$ as follows:

$$X(t) \triangleq \begin{cases} 1 & \text{for } t = U \\ 0, & \text{else,} \end{cases}$$

and

$$Y(t) \triangleq 0 \quad \text{for all } t.$$

Then $Y(t)$ and $X(t)$ will have the same finite-order distributions, yet obviously the probability of the following two events is not the same:

$$\{X(t) \leq 0.5 \text{ for all } t\}$$

and

$$\{Y(t) \leq 0.5 \text{ for all } t\}.$$

To show that $Y(t)$ and $X(t)$ have the same nth-order pdf's, find the conditional nth-order pdf of X given $U = u$, then integrate out the conditioning on U. We leave this as an exercise to the reader.

The problem in Example 9.1-4 is that the complementary event $\{X(t) > 0.5\}$ for some $t \in [0, 1]\}$ involves an *uncountable* number of random variables. Yet the statistical determination and the extended additivity Axiom 4 (see Section 8.1) only allow us to evaluate probabilities corresponding to *countable* numbers of random variables. In what follows, we will generally assume that we always have a process "continuous enough" that the family of finite-order distribution functions suffices to determine the process for all time.† Such processes are called *separable*. The random process $X(t)$ of the above example is obviously not separable.

As in the case of random sequences, the moment functions play an important role in practical applications. The *mean function*, denoted by $\mu_X(t)$, is given as

$$\mu_X(t) \triangleq E[X(t)], \qquad -\infty < t < +\infty. \tag{9.1-1}$$

Similarly the *correlation function* is defined as the expected value of the conjugate product,

$$R_{XX}(t_1, t_2) \triangleq E[X(t_1)X^*(t_2)], \qquad -\infty < t_1, t_2 < +\infty. \tag{9.1-2}$$

The *covariance function* is defined as the expected value of the conjugate product of the *centered process* $X_c(t) \triangleq X(t) - \mu_X(t)$ at times t_1 and t_2:

$$\begin{aligned} K_{XX}(t_1, t_2) &\triangleq E[X_c(t_1)X_c^*(t_2)] \\ &\triangleq E[(X(t_1) - \mu_X(t_1))(X(t_2) - \mu_X(t_2))^*]. \end{aligned} \tag{9.1-3}$$

†An exception is white noise to be introduced in Section 9.3.

Clearly these three functions are not unrelated and in fact we have,

$$K_{XX}(t_1,t_2) = R_{XX}(t_1,t_2) - \mu_X(t_1)\mu_X^*(t_2). \tag{9.1-4}$$

We also define the *variance function* as $\sigma_X^2(t) \triangleq K_{XX}(t,t) = E\left[|X_c(t)|^2\right]$, and the *power function* $R_{XX}(t,t) = E\left[|X(t)|^2\right]$.

Example 9.1-5

(*more on random sinewave*) Consider the random process

$$X(t) = A\sin(\omega_0 t + \Theta),$$

where A and Θ are independent, real-valued random variables and Θ is uniformly distributed over $[-\pi, +\pi]$. For this sinusoidal random process, we will find the mean function $\mu_X(t)$ and correlation function $R_{XX}(t_1,t_2)$. First

$$\begin{aligned}
\mu_X(t) &= E[A\sin(\omega_0 t + \Theta)] \\
&= E[A]E[\sin(\omega_0 t + \Theta)] \\
&= \mu_A \cdot \frac{1}{2\pi}\int_{-\pi}^{+\pi} \sin(\omega_0 t + \theta)d\theta \\
&= \mu_A \cdot 0 = 0.
\end{aligned}$$

Then for the correlation,

$$\begin{aligned}
R_{XX}(t_1,t_2) &= E[X(t_1)X^*(t_2)] \\
&= E[A^2\sin(\omega_0 t_1 + \Theta)\sin(\omega_0 t_2 + \Theta)] \\
&= E[A^2]E[\sin(\omega_0 t_1 + \Theta)\sin(\omega_0 t_2 + \Theta)].
\end{aligned}$$

Now, the second factor can be rewritten as

$$\tfrac{1}{2}\{E[\cos(\omega_0(t_1 - t_2))] - E[\cos(\omega_0(t_1 + t_2) + 2\Theta)]\} \tag{9.1-5}$$

by applying the trigonometric identity

$$\sin(B)\sin(C) = \tfrac{1}{2}\{\cos(B-C) - \cos(B+C)\},$$

and bringing the expectation operator inside. Then, since Θ is uniformly distributed over $[-\pi, +\pi]$, the integral arising from the second expectation in Equation 9.1-5 is zero, and we finally obtain

$$R_{XX}(t_1,t_2) = \tfrac{1}{2}E[A^2]\cos\omega_0(t_1 - t_2).$$

We note that $\mu_X(t) = 0$ (a constant) and $R_{XX}(t_1,t_2)$ depends only on $t_1 - t_2$. Such processes will be classified as *wide-sense stationary* in Section 9.4.

As in the discrete-time case, the correlation and covariance functions are Hermitian symmetric, that is,

$$R_{XX}(t_1, t_2) = R^*_{XX}(t_2, t_1),$$

$$K_{XX}(t_1, t_2) = K^*_{XX}(t_2, t_1),$$

which directly follow from the linearity of the expectation operator E.

If we sample the random process at N times $t_1, t_2, \ldots, t_N$, we form a *random vector*. We have already seen that the correlation or covariance matrix of a random vector must be positive semidefinite (cf. Chapter 5). This, then, imposes certain requirements on the respective correlation and covariance function of the random process. Specifically, every correlation (covariance) matrix that can be formed from a correlation (covariance) function must be positive semidefinite. We next define *positive semidefinite* functions.

Definition 9.1-3 The two-dimensional function $g(t, s)$ is *positive semidefinite* if for all $N > 0$, and all $t_1 < t_2 < \ldots < t_N$, and for all complex constants $a_1, a_2, \ldots, a_N$, we have

$$\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j^* g(t_i, t_j) \geq 0. \quad \blacksquare$$

Using this definition, we can thus say that all correlation and covariance functions must be positive semidefinite. Later we will see that this necessary condition is also sufficient. Although positive semidefiniteness is an important constraint, it is difficult to apply this condition in a test of the legitimacy of a proposed correlation function.

Another fundamental property of correlation and covariance functions is *diagonal dominance*,

$$|R_{XX}(t, s)| \leq \sqrt{R_{XX}(t, t)R_{XX}(s, s)} \quad \text{for all } t, s,$$

which follows from the Cauchy–Schwarz inequality (cf. Equation 4.3-17). Diagonal dominance is implied by positive semidefiniteness but is a much weaker condition.

9.2 SOME IMPORTANT RANDOM PROCESSES

In this section we introduce several important random processes. We start with the asynchronous binary signaling (ABS) process and the random telegraph signal (RTS). We continue with the Poisson counting process; the phase-shift keying (PSK) random process, an example of digital modulation; the Wiener process, which is obtained as a continuous limit of a random walk sequence; and lastly introduce the broad class of Markov processes.

Asynchronous Binary Signaling

A sample function of the asynchronous binary signaling (ABS) process (important for digital modulation and computers) is shown in Figure 9.2-1. Each pulse has width T with the

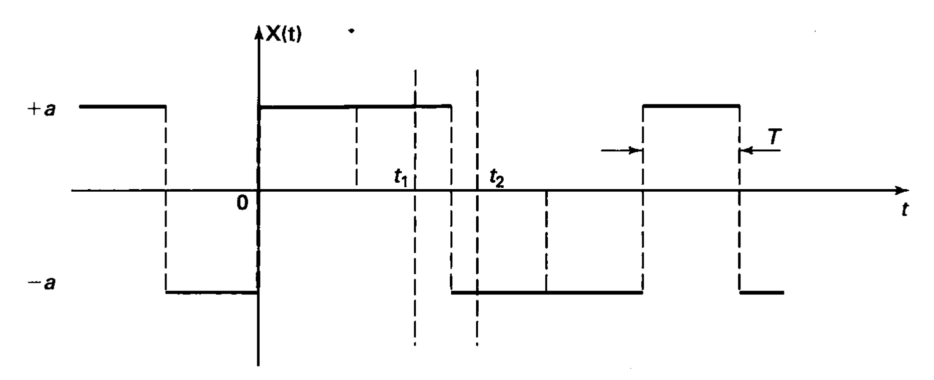


Figure 9.2-1 Sample function realization of the asynchronous binary signaling (ABS) process. (Plotted for $D = 0$.)

random variable X_n indicating the height of the nth pulse, taking on values $\pm a$ with equal probability.

The sequence is *asynchronous* because the start time of the nth pulse or, equivalently, the displacement D of the 0th pulse is a uniform random variable $U(-\frac{T}{2}, \frac{T}{2})$. For $|t_2 - t_1| < T$, the sampling instant t_2 could be on the same pulse containing the sampling instant t_1 or on a different pulse.

The ABS process can thus be described mathematically by

$$X(t) = \sum_n X_n \, w\left[\frac{t - D - nT}{T}\right],$$

where the pulse (rectangular window) function $w(t)$ is defined as

$$w(t) \triangleq \begin{cases} 1 & \text{for } |t| \le \frac{1}{2} \\ 0 & \text{else.} \end{cases}$$

The correlation function for this real-valued process is given as

$$\begin{aligned} R_{XX}(t_1, t_2) &= E[X(t_1)X(t_2)] \\ &= E\left[\sum_n \sum_l X_n X_l \, w\left(\frac{t_1 - D - nT}{T}\right) \, w\left(\frac{t_2 - D - lT}{T}\right)\right]. \end{aligned}$$

In the ABS process it is assumed the levels of different pulses are independent random variables and that these, in turn, are independent of the random displacement D. Since $E[X_n X_l] = E[X_n]E[X_l]$ for $n \neq l$ and $E[X_n^2] = a^2$, we obtain

$$\begin{aligned} R_{XX}(t_1, t_2) &= a^2 \sum_n E\left[w\left(\frac{t_1 - D - nT}{T}\right) \, w\left(\frac{t_2 - D - nT}{T}\right)\right] \\ &+ \sum_{n \neq l} \sum_l E[X_n]E[X_l] \, E\left[w\left(\frac{t_1 - D - nT}{T}\right) \, w\left(\frac{t_2 - D - lT}{T}\right)\right]. \end{aligned}$$

Now, the second term on the right, the one involving the $n \neq l$ products, is zero because $E[X_n] = E[X_l] = 0$. Also

$$\begin{aligned}
&\sum_n E\left[w\left(\frac{t_1 - D - nT}{T}\right) w\left(\frac{t_2 - D - nT}{T}\right)\right] \\
&= \sum_n \frac{1}{T}\int_{t_2-\frac{T}{2}-nT}^{t_2+\frac{T}{2}-nT} w\left(\frac{\alpha}{T}\right) w\left(\frac{\alpha - (t_2 - t_1)}{T}\right) d\alpha \\
&= \left(1 - \frac{(t_2 - t_1)}{T}\right) w\left(\frac{t_2 - t_1}{2T}\right) \quad \text{for } t_2 > t_1.
\end{aligned}$$

More generally, and for $\tau \triangleq t_2 - t_1 \lesseqgtr 0$, we can write that

$$R_{XX}(\tau) = a^2\left(1 - \frac{|\tau|}{T}\right) w\left(\frac{\tau}{2T}\right) \tag{9.2-1}$$

since $w(|\tau|) = w(\tau)$.

Equation 9.2-1 is directly extended to the case of equiprobable transitions between two arbitrary levels, say a and b. The required modification is

$$R_{XX}(\tau) = \frac{1}{4}(a-b)^2\left(1 - \frac{|\tau|}{T}\right) w\left(\frac{\tau}{2T}\right) + \left(\frac{a+b}{2}\right)^2.$$

We leave the derivation of this result as an exercise for the reader. In Figure 9.2-2 we show the ABS correlation function $R_{XX}(\tau)$ for $a = 1, b = 0$, and $T = 1$.

Poisson Counting Process

Let the process $N(t)$ represent the total number of counts (arrivals) up to time t. Then we can write

$$N(t) \triangleq \sum_{n=1}^{\infty} u(t - T[n]),$$

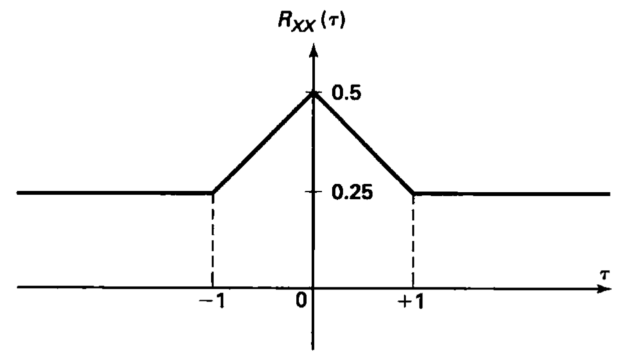


Figure 9.2-2 Autocorrelation function of ABS random process for $a = 1, b = 0$ and $T = 1$.

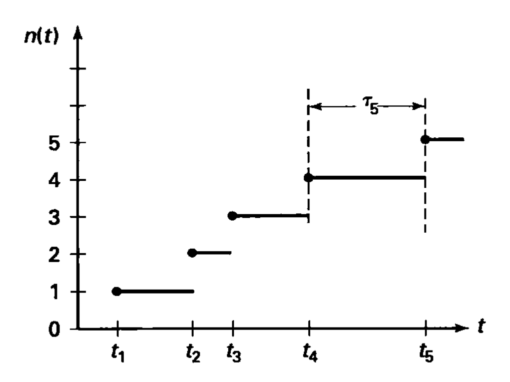


Figure 9.2-3 A sample function of the Poisson process running on $[0, \infty)$.

where $u(t)$ is the unit-step function and $T[n]$, *the time to the nth arrival,* is the random sequence of times considered in Example 8.1-11. There we showed that the $T[n]$ obeyed the nonstationary first-order Erlang density,

$$f_T(t;n) = \frac{(\lambda t)^{n-1}}{(n-1)!}\lambda e^{-\lambda t}u(t), \qquad n \geq 0, \tag{9.2-2}$$

which was obtained as an n-fold convolution of exponential pdf's. A typical sample function is shown in Figure 9.2-3, where $T[n] = t_n$ and $\tau[n] = \tau_n$. Note that the time between the arrivals,

$$\tau[n] \triangleq T[n] - T[n-1],$$

the *interarrival times*, are jointly independent and identically distributed, having the exponential pdf,

$$f_\tau(t) = \lambda e^{-\lambda t}u(t),$$

as in Example 8.1-11. Thus, $T[n]$ denotes the total time until the nth arrival if we begin counting at the reference time $t = 0$.

Now by the construction involving the unit-step function, the value $N(t)$ is the number of arrivals up to and including time t, so

$$P[N(t) = n] = P[T[n] \leq t, T[n+1] > t],$$

because the only way that $N(t)$ can equal n is if the random variable $T[n]$ is less than or equal to t and the random variable $T[n+1]$ is greater than t. If we bring in the independent interarrival times, we can re-express this probability as

$$P[T[n] \leq t, \tau[n+1] > t - T[n]],$$

which can be easily calculated using the statistical independence of the arrival time, $T[n]$ and the interarrival time $\tau[n+1]$ as follows:

$$\int_0^t f_T(\alpha;n)\left[\int_{t-\alpha}^{\infty} f_\tau(\beta)d\beta\right]d\alpha = \int_0^t \frac{\lambda^n \alpha^{n-1}e^{-\lambda\alpha}}{(n-1)!}\left(\int_{t-\alpha}^{\infty}\lambda e^{-\lambda\beta}d\beta\right)d\alpha \cdot u(t)$$
$$= \left(\int_0^t \alpha^{n-1}d\alpha\right)\lambda^n e^{-\lambda t}/(n-1)!\ u(t),$$

or, with $P_N(n;t) \triangleq P\{N(t)=n\}$,

$$P_N(n;t) = \frac{(\lambda t)^n}{n!}e^{-\lambda t}u(t) \quad \text{for } t \geq 0, \qquad n \geq 0. \tag{9.2-3}$$

We have thus arrived at the PMF of the *Poisson counting process* and we note that it's equal to that of a Poisson random variable (cf. Equation 2.5-13, see also Equation 1.10-5) with mean $\mu = \lambda t$ is

$$E[N(t)] = \lambda t. \tag{9.2-4}$$

We call λ the *mean arrival rate* (also sometimes called *intensity*). It is intuitively satisfying that the average value of the process at time t is the mean arrival rate λ multiplied by the length of the time interval $(0,t]$. We leave it as an exercise for the reader to consider why this is so.

Since the random sequence $T[n]$ has independent increments (cf. Definition 8.1-4) and the unit-step function used in the definition of the Poisson process is causal, it seems reasonable that the Poisson process $N(t)$ would also have independent increments. However, this result is not clear because one of the jointly independent interarrival times $\tau[n]$ may be partially in two disjoint intervals, hence causing a dependency in neighboring increments. Nevertheless, using the *memoryless property* of the exponential pdf (see Problem 9.8), one can show that the independent-increments property does hold for the Poisson process.

Using independent increments we can evaluate the PMF of the increment in the Poisson counting process over an interval (t_a, t_b) as

$$P[N(t_b) - N(t_a) = n] = \frac{[\lambda(t_b - t_a)]^n}{n!}e^{-\lambda(t_b - t_a)}u(n), \tag{9.2-5}$$

where we have used the fact that the interarrival sequence is stationary, that is, that λ is a constant. We formalize this somewhat in the following definition.

Definition 9.2-1 A random process has *independent increments* when the set of n random variables,

$$X(t_1), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1}),$$

are jointly independent for all $t_1 < t_2 < \ldots < t_n$ and for all $n \geq 1$. ■

This just says that the increments are statistically independent when the corresponding intervals do not overlap. Just as in the random sequence case, the independent-increment

property makes it easy to get the higher-order distributions. For example, in the case at hand, the Poisson counting process, we can write for $t_2 > t_1$,

$$\begin{aligned} P_N(n_1, n_2; t_1, t_2) &= P[N(t_1) = n_1]\ P[N(t_2) - N(t_1) = n_2 - n_1] \\ &= \frac{(\lambda t_1)^{n_1}}{n_1!} e^{-\lambda t_1} \frac{[\lambda(t_2 - t_1)]^{n_2 - n_1}}{(n_2 - n_1)!} e^{-\lambda(t_2 - t_1)} u(n_1) u(n_2 - n_1), \end{aligned}$$

which simplifies to

$$P_N(n_1, n_2; t_1, t_2) = \frac{\lambda^{n_2} t_1^{n_1} (t_2 - t_1)^{n_2 - n_1}}{n_1!(n_2 - n_1)!} e^{-\lambda t_2} u(n_1) u(n_2 - n_1), \qquad 0 \leq t_1 < t_2.$$

See also Problem 1.54. Using the independent-increments property we can formulate the following alternative definition of a Poisson counting process.

Definition 9.2-2 A Poisson counting process is the independent-increments process whose increments are Poisson distributed as in Equation 9.2-5. ■

Concerning the moment function of the Poisson process, the first-order moment has been shown to be λt. This is the mean function of the process. Letting $t_2 \geq t_1$, we can calculate the correlation function using the independent-increments property as

$$\begin{aligned} E[N(t_2)N(t_1)] &= E[(N(t_1) + [N(t_2) - N(t_1)])N(t_1)] \\ &= E[N^2(t_1)] + E[N(t_2) - N(t_1)]E[N(t_1)] \\ &= \lambda t_1 + \lambda^2 t_1^2 + \lambda(t_2 - t_1)\lambda t_1 \\ &= \lambda t_1 + \lambda^2 t_1 t_2. \end{aligned}$$

If $t_2 < t_1$, we merely interchange t_1 and t_2 in the preceding formula. Thus the general result for all t_1 and t_2 is

$$\begin{aligned} R_{NN}(t_1, t_2) &= E[N(t_1)N(t_2)] \\ &= \lambda \min(t_1, t_2) + \lambda^2 t_1 t_2. \end{aligned} \tag{9.2-6}$$

If we evaluate the covariance using Equations 9.2-4 and 9.2-6 we obtain

$$K_{NN}(t_1, t_2) = \lambda \min(t_1, t_2). \tag{9.2-7}$$

We thus see that the variance of the process is equal to λt and is the same as its mean, a property inherited from the Poisson random variable. Also we see that the covariance depends only on the earlier of the two times involved. The reason for this is seen by writing $N(t)$ as the value at an earlier time plus an increment, and then noting that the independence of this increment and $N(t)$ at the earlier time implies that the covariance between them must be zero. Thus, the covariance of this independent-increments process is just the variance of the process at the earlier of the two times.

Example 9.2-1

(*radioactivity monitor*) In radioactivity monitoring, the particle-counting process can often be adequately modeled as Poisson. Let the counter start to monitor at some arbitrary time t

and then count for T_0 seconds. If the count is above a threshold, say N_0, an alarm will be sounded. Assuming the arrival rate to be λ, we want to know the probability that the alarm will not sound when radioactive material is present.

Since the process is Poisson, we know it has independent increments that satisfy the Poisson distribution. Thus the count ΔN in the interval $(t, t+T_0]$, that is, $\Delta N \triangleq N(t+T_0) - N(t)$, is Poisson distributed with mean λT_0 independent of t. The probability of N_0 or fewer counts is thus

$$P[\Delta N \leq N_0] = \sum_{k=0}^{N_0} \frac{(\lambda T_0)^k}{k!} e^{-\lambda T_0}.$$

If N_0 is small we can calculate the sum directly. If $\lambda T_0 >> 1$, we can use the Gaussian approximation (Equation 1.11-9) to the Poisson distribution.

Example 9.2-2

(*sum of two independent Poisson processes*) Let $N_1(t)$ be a Poisson counting process with rate λ_1. Let $N_2(t)$ be a second Poisson counting process with rate λ_2, where N_2 is independent of N_1. The sum of the two processes, $N(t) \triangleq N_1(t) + N_2(t)$, could model the total number of failures of two separate machines, whose failure rates are λ_1 and λ_2, respectively. It is a remarkable fact that $N(t)$ is also a Poisson counting process with rate $\lambda = \lambda_1 + \lambda_2$.

To see this we use Definition 9.2-2 of the Poisson counting process and verify these conditions for $N(t)$. First, it is clear with a little reflection that the sum of two independent-increments processes will also be an independent-increments process *if* the processes are jointly independent. Second, for any increment $N(t_b) - N(t_a)$ with $t_b > t_a$, we can write

$$N(t_b) - N(t_a) = N_1(t_b) - N_1(t_a) + N_2(t_b) - N_2(t_a).$$

Thus the increment in N is the sum of two corresponding increments in N_1 and N_2. The desired result then follows from the fact that the sum of two independent Poisson random variables is also Poisson distributed with parameter equal to the sum of the two parameters (cf. Example 3.3-8). Thus the parameter of the increment in $N(t)$ is

$$\lambda_1(t_b - t_a) + \lambda_2(t_b - t_a) = (\lambda_1 + \lambda_2)(t_b - t_a)$$

as desired.

The Poisson counting process $N(t)$ can be generalized in several ways. We can let the arrival rate be a function of time. The arrival rate $\lambda(t)$ must satisfy $\lambda(t) \geq 0$. The average value of the resulting *nonuniform Poisson counting process* then becomes

$$\mu_X(t) = \int_0^t \lambda(\tau)d\tau, \qquad t \geq 0. \tag{9.2-8}$$

The increments then become independent Poisson distributed with increment means determined by this time-varying mean function. Another possible generalization is to two-dimensional or spatial Poisson processes that are used to model photon arrival at an image sensor, defects on semiconductor wafers, etc.

Alternative Derivation of Poisson Process

It may be interesting to rederive the Poisson counting process from the elementary properties of random points in time listed in Chapter 1, Section 1.10. They are repeated here in a notation consistent with that used in this chapter. For Δt small:

(1) $P_N(1;t,t+\Delta t) = \lambda(t)\Delta t + o(\Delta t)$.
(2) $P_N(k;t,t+\Delta t) = o(\Delta t), \qquad k > 1$.
(3) Events in nonoverlapping time intervals are statistically independent.

Here the notation $o(\Delta t)$, read "little oh," denotes any quantity that goes to zero at a faster than linear rate in such a way that

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0,$$

and $P_N(k;t,t+\Delta t) = P[N(t+\Delta t) - N(t) = k]$.

We note that property (3) is just the independent-increments property for the counting process $N(t)$ which counts the number of events occurring in $(0,t]$.

We can compute the probability $P_N(k;t,t+\tau)$ of k events in $(t,t+\tau)$ as follows. Consider $P_N(k;t,t+\tau+\Delta t)$; if Δt is very small, then in view of properties (1) and (2) there are only the following two possibilities for getting k events in $(t,t+\tau+\Delta t)$:

$$E_1 = \{k \text{ in } (t,t+\tau) \text{ and } 0 \text{ in } (t+\tau,t+\tau+\Delta t)\} \qquad \text{or}$$

$$E_2 = \{k-1 \text{ in } (t,t+\tau) \text{ and } 1 \text{ in } (t+\tau,t+\tau+\Delta t)\}.$$

Since events E_1 and E_2 are disjoint events, their probabilities add and we can write

$$\begin{aligned} P_N(k;t,t+\tau+\Delta t) &= P_N(k;t,t+\tau)P_N(0;t+\tau,t+\tau+\Delta t) \\ &\quad + P_N(k-1;t,t+\tau)P_N(1;t+\tau,t+\tau+\Delta t) \\ &= P_N(k;t,t+\tau)[1-\lambda(t+\tau)\Delta t] \\ &\quad + P_N(k-1;t,t+\tau)\lambda(t+\tau)\Delta t. \end{aligned}$$

If we rearrange terms, divide by Δt, and take limits, we obtain the linear differential equations (LDEs),

$$\frac{dP_N(k;t,t+\tau)}{d\tau} = \lambda(t+\tau)[P_N(k-1;t,t+\tau) - P_N(k;t,t+\tau)].$$

Thus, we obtain a set of recursive first-order differential equations from which we can solve for $P_N(k;t,t+\tau), k = 0,1,\ldots$. We set $P_N(-1;t,t+\tau) = 0$, since this is the probability of the impossible event. Also, to shorten our notation, we temporarily write $P_N(k) \triangleq P_N(k;t,t+\tau)$; thus the dependences on t and τ are submerged but of course are still there.

When $k = 0$,

$$\frac{dP_N(0)}{d\tau} = -\lambda(t+\tau)P_N(0).$$

This is a simple first-order, homogeneous differential equation for which the solution is

$$P_N(0) = C\exp\left[-\int_t^{t+\tau} \lambda(\xi)d\xi\right].$$

Since $P_N(0; t, t) = 1, C = 1$ and

$$P_N(0) = \exp\left[-\int_t^{t+\tau} \lambda(\xi)d\xi\right].$$

Let us define μ by

$$\mu \triangleq \int_t^{t+\tau} \lambda(\xi)d\xi.$$

Then

$$P_N(0) = e^{-\mu}.$$

When $k = 1$, the differential equation is now

$$\begin{aligned}\frac{dP_N(1)}{d\tau} + \lambda(t+\tau)P_N(1) &= \lambda(t+\tau)P_N(0) \\ &= \lambda(t+\tau)e^{-\mu}.\end{aligned} \tag{9.2-9}$$

This elementary first-order, inhomogeneous equation has a solution that is the sum of the homogeneous and particular solutions. For the homogeneous solution, P_h, we already know from the $k = 0$ case that

$$P_h = C_2 e^{-\mu}.$$

For the particular solution P_p we use the method of *variation of parameters* to assume that

$$P_p = v(t+\tau)e^{-\mu},$$

where $v(t+\tau)$ is to be determined. By substituting this equation into Equation 9.2-9 we readily find that

$$P_p = \mu e^{-\mu}.$$

The complete solution is $P_N(1) = P_h + P_p$. Since $P_N(1; t, t) = 0$, we obtain $C_2 = 0$ and thus

$$P_N(1) = \mu e^{-\mu}.$$

General case. The LDE in the general case is

$$\frac{dP_N(k)}{d\tau} + \lambda(t+\tau)P_N(k) = \lambda(t+\tau)P_N(k-1)$$

and, proceeding by induction, we find that

$$P_N(k) = \frac{\mu^k}{k!}e^{-\mu} \qquad k = 0, 1, \ldots$$

which is the key result. Recalling the definition of μ, we can write

$$P_N(k; t, t+\tau) = \frac{1}{k!}\left[\int_t^{t+\tau} \lambda(\xi)d\xi\right]^k \exp\left[-\int_t^{t+\tau} \lambda(\xi)d\xi\right]. \tag{9.2-10}$$

We thus obtain the nonuniform Poisson counting process.

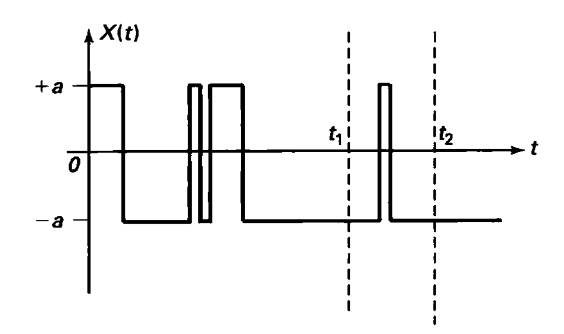


Figure 9.2-4 Sample function of the random telegraph signal.

Another way to generalize the Poisson process is to use a different pdf for the independent interarrival times. With a nonexponential density, the more general process is called a *renewal process* [9-2]. The word "renewal" can be related to the interpretation of the arrival times as the failure times of certain equipment; thus the value of the counting process $N(t)$ models the number of renewals that have had to be made up to the present time.

Random Telegraph Signal

When all the information in a random waveform is contained in the *zero crossings*, a so-called "hard clipper" is often used to generate a simpler yet equivalent two-level waveform that is free of unwanted random amplitude variation. A special case is when the number of zero crossings in a time interval follows the Poisson law, and the resulting random process is called the *random telegraph signal* (RTS). A sample function of the RTS is shown in Figure 9.2-4.

We construct the RTS on $t \geq 0$ as follows: Let $X(0) = \pm a$ with equal probability. Then take the Poisson arrival time sequence $T[n]$ of Chapter 8 and use it to switch the level of the RTS; that is, at $T[1]$ switch the sign of $X(t)$, and then at $T[2]$, and so forth. Clearly from the symmetry and the fact that the interarrival times $\tau[n]$ are stationary and form an independent random sequence, we must have that $\mu_X(t) = 0$ and that the first-order PMF $P_X(a) = P_X(-a) = 1/2$. Next let $t_2 > t_1 > 0$, and consider the second-order PMF $P_X(x_1, x_2) \triangleq P[X(t_1) = x_1, X(t_2) = x_2]$ along with $P_X(x_2 \mid x_1) \triangleq P[X(t_2) = x_2 \mid X(t_1) = x_1]$. Then we can write the correlation function as

$$\begin{aligned} R_{XX}(t_1, t_2) &= E[X(t_1)X(t_2)] \\ &= a^2 P_X(a, a) + (-a)^2 P_X(-a, -a) + a(-a)P_X(a, -a) - a(a)P_X(-a, a) \\ &= \frac{1}{2}a^2(P_X(a|a) + P_X(-a|-a) - P_X(-a|a) - P_X(a|-a)), \end{aligned}$$

since $P_X(a) = P_X(-a) = 1/2$. But $P_X(-a|-a) = P_X(a|a)$ is just the probability of an even number of zero crossings in the time interval $(t_1, t_2]$, while $P_X(-a|a) = P_X(a|-a)$ is the probability of an odd number of crossings of 0. Hence, writing the average number of transitions per unit time as λ, and substituting $\tau \triangleq t_2 - t_1$, we get

$$R_{XX}(t_1,t_2) = a^2\left(\sum_{\text{even } k\geq 0} e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!} - \sum_{\text{odd } k\geq 0} e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!}\right) = a^2 e^{-\lambda\tau}\sum_{\text{all } k\geq 0}(-1)^k\frac{(\lambda\tau)^k}{k!},$$

where we have combined the two sums by making use of the function $(-1)^k$, since $(-1)^k = 1$ for k even and $(-1)^k = -1$ for k odd. Thus we now have

$$R_{XX}(t_1,t_2) = a^2 e^{-\lambda\tau}\sum_{\text{all } k\geq 0}\frac{(-\lambda\tau)^k}{k!} = a^2 e^{-2\lambda\tau}$$

for the case when $\tau > 0$. Since the correlation function of a real-valued process must be symmetric, we have $R_{XX}(t_1,t_2) = R_{XX}(t_2,t_1)$, so that when $\tau \leq 0$, we can substitute $-\tau$ into the above equation to get $R_{XX}(t_1,t_2) = a^2 e^{+2\lambda\tau}$. Thus overall we have, valid for all interval lengths τ,

$$R_{XX}(t_1,t_2) = a^2 e^{-2\lambda|\tau|}.$$

A plot of this correlation function is shown in Figure 9.2-5.

Digital Modulation Using Phase-Shift Keying

Digital computers generate many binary sequences (data) to be communicated to other digital computers. Often this involves some kind of modulation. Binary modulation methods frequency-shift these data to a region of the electromagnetic spectrum which is well suited to the transmission media, for example, a telephone line. A basic method for modulating binary data is phase-shift keying (PSK). In this method binary data, modeled by the random sequence $B[n]$, are mapped bit-by-bit into a phase-angle sequence $\Theta[n]$, which is used to modulate a *carrier signal* $\cos(2\pi f_c t)$.

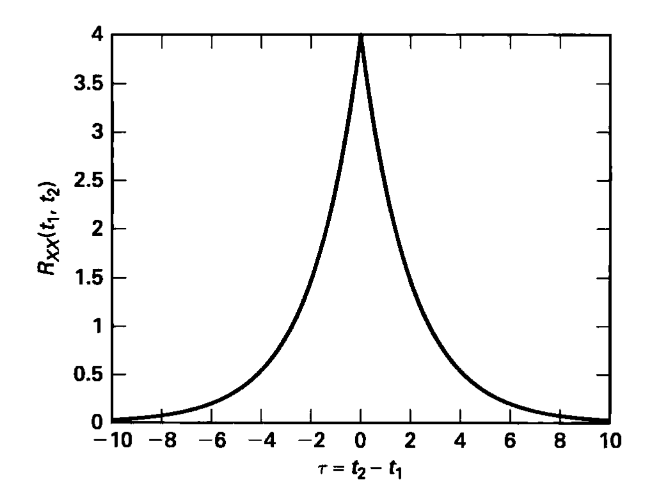


Figure 9.2-5 The symmetric exponential correlation function of an RTS process ($a = 2.0, \lambda = 0.25$).

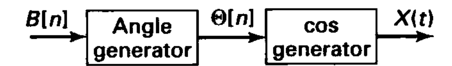


Figure 9.2-6 System for PSK modulation of Bernoulli random sequence $B[n]$.

Specifically let $B[n]$ be a Bernoulli random sequence taking on the values 0 and 1 with equal probability. Then define the random phase sequence $\Theta[n]$ as follows:

$$\Theta[n] \triangleq \begin{cases} +\pi/2 & \text{if } B[n] = 1, \\ -\pi/2 & \text{if } B[n] = 0. \end{cases}$$

Using $\Theta_a(t)$ to denote the analog angle process, we define

$$\Theta_a(t) \triangleq \Theta[k] \quad \text{for } kT \leq t < (k+1)T,$$

and construct the modulated signal as

$$X(t) = \cos(2\pi f_c t + \Theta_a(t)). \tag{9.2-11}$$

Here T is a constant time for the transmission of one bit. Normally, T is chosen to be a multiple of $1/f_c$ so that there are an integral number of carrier cycles per bit time T. The reciprocal of T is called the message or *baud rate.* The overall modulator is shown in Figure 9.2-6. The process $X(t)$ is the PSK process.

Our goal here is to evaluate the mean function and correlation function of the random PSK process. To help in the calculation we define two basis functions,

$$s_I(t) \triangleq \begin{cases} \cos(2\pi f_c t) & 0 \leq t \leq T \\ 0 & \text{else,} \end{cases}$$

and

$$s_Q(t) \triangleq \begin{cases} \sin(2\pi f_c t) & 0 \leq t \leq T \\ 0 & \text{else,} \end{cases}$$

which together with Equation 9.2-11 imply

$$\begin{aligned} \cos[2\pi f_c t + \Theta_a(t)] &= \cos(\Theta_a(t)) \cos 2\pi f_c t - \sin(\Theta_a(t)) \sin 2\pi f_c t \\ &= \sum_{k=-\infty}^{+\infty} \cos(\Theta[k]) s_I(t - kT) - \sum_{k=-\infty}^{+\infty} \sin(\Theta[k]) s_Q(t - kT), \end{aligned} \tag{9.2-12}$$

by use of the sum of angles formula for cosines.

The mean of $X(t)$ can then be obtained in terms of the means of the random sequences $\cos(\Theta[n])$ and $\sin(\Theta[n])$. Because of the definition of $\Theta[n]$, in this particular case $\cos(\Theta[n]) = 0$ and $\sin(\Theta[n]) = \pm 1$ with equal probability so that mean of $X(t)$ is zero, that is, $\mu_X(t) = 0$.

Using Equation 9.2-12 we can calculate the correlation function

$$R_{XX}(t_1, t_2) = \sum_{k,l} E\{\sin\Theta[k]\sin\Theta[l]\} s_Q(t_1 - kT)s_Q(t_2 - lT),$$

which involves the correlation function of the random sequence $\sin(\Theta[n])$,

$$R_{\sin\Theta,\sin\Theta}[k,l] = \delta[k-l].$$

Thus the overall correlation function then becomes

$$R_{XX}(t_1, t_2) = \sum_{k=-\infty}^{+\infty} s_Q(t_1 - kT)s_Q(t_2 - kT). \tag{9.2-13}$$

Since the support of s_Q is only of width T, there is no overlap in (t_1, t_2) between product terms in Equation 9.2-13. So for any fixed (t_1, t_2), only one of the product terms in the sum can be nonzero. Also if t_1 and t_2 are not in the same period, then this term is zero also. More elegantly, using the notation,

$$(t) \triangleq t \bmod T \quad \text{and} \quad \lfloor t/T \rfloor \triangleq \text{integer part } (t/T),$$

we can write that

$$R_{XX}(t_1, t_2) = \begin{cases} s_Q((t_1))s_Q((t_2)) & \text{for } \lfloor t_1/T \rfloor = \lfloor t_2/T \rfloor \\ 0 & \text{else.} \end{cases}$$

In particular for $0 \le t_1 \le T$ and $0 \le t_2 \le T$, we have

$$R_{XX}(t_1, t_2) = s_Q(t_1)s_Q(t_2).$$

Wiener Process or Brownian Motion

In Chapter 8 we considered a random sequence $X[n]$ called the random walk in Example 8.1-13. Here we construct an analogous *random process* that is piecewise constant for intervals of length T as follows:

$$X_T(t) \triangleq \sum_{k=1}^{\infty} W[k]u(t - kT),$$

where

$$W[k] \triangleq \begin{cases} +s & \text{with } p = 0.5 \\ -s & \text{with } p = 0.5 \end{cases}$$

and $u(t)$ is the continuous unit step function.
Then $X_T(nT) = X[n]$ the random-walk sequence, since

$$X_T(nT) = \sum_{k=1}^{n} W[k] = X[n].$$

Hence we can evaluate the PMFs and moments of this random process by employing the known results for the corresponding random-walk sequence. Now the Wiener† process,

†After Norbert Wiener, American mathematician (1894–1964), a pioneer in communication and estimation theories.

sometimes also called *Wiener–Levy* or Brownian motion, is the process whose distribution is obtained as a limiting form of the distribution of the above piecewise constant process as the interval T shrinks to zero. We let s, the jump size, and the interval T shrink to zero in a precise way to obtain a *continuous random process* in the limit, that is, a process whose sample functions are continuous functions of time. In letting s and T tend to zero we must be careful to make sure that the limit of the variance stays finite and nonzero. The resulting Wiener process will inherit the independent-increments property.

The original motivation for the Wiener process was to develop a model for the chaotic random motion of gas molecules. Modeling the basic discrete collisions with a random walk, one then finds the asymptotic process when an infinite (very large) number of molecules interact on an infinitesimal (very small) time scale.

As in Example 8.1-13, we let n be the number of trials, k be the number of successes, and $n-k$ be the number of failures. Also $r \triangleq k-(n-k) = 2k-n$ denotes the excess number of successes over failures. Then $2k = n+r$ or $k = (n+r)/2$ and must be an integer; you cannot have 2.5 "successes." Thus, $n+r$ must be even and the probability that $X_T(nT) = rs$ is the probability that there are $0.5(n+r)$ successes $(+s)$ and $0.5(n-r)$ failures $(-s)$ out of a total of n trials. Thus by the binomial PMF,

$$P[X_T(nT) = rs] = \binom{n}{\frac{n+r}{2}} 2^{-n} \qquad \text{for } n+r \text{ even.}$$

If $n+r$ is odd, then $X_T(nT)$ cannot equal rs.

The mean and variance can be most easily calculated by noting that the random variable $X[n]$ is the sum of n independent Bernoulli random variables defined in Section 8.1. Thus

$$E[X_T(nT)] = 0$$

and

$$E[X_T^2(nT)] = ns^2.$$

On expressing the variance in terms of $t = nT$, we have

$$\text{Var}[X_T(t)] = E[X_T^2(nT)] = t\frac{s^2}{T}.$$

Thus we need s^2 proportional to T to get an interesting limiting distribution.† We set $s^2 = \alpha T$, where $\alpha > 0$. Now as T goes to zero we keep the variance constant at αt. Also, by an elementary application of the Central Limit theorem (cf. Section 4.7), we get a limiting Gaussian distribution. We take the limiting random process (convergence in the distribution sense) to be an independent-increments process since all the above random-walk processes had independent increments for all T, no matter how small. Hence we arrive at the following specification for the limiting process, which is termed the *Wiener process*:

$$\mu_X(t) = 0, \qquad \text{Var}[X(t)] = \alpha t$$

†The physical implication of having s^2 proportional to T is that if we take $v \triangleq s/T$ as the speed of the particle, then the particle speed goes to infinity as the displacement s goes to zero such as to keep the product of the two constant.

and

$$f_X(x;t) = \frac{1}{\sqrt{2\pi\alpha t}} \exp\left(-\frac{x^2}{2\alpha t}\right), \qquad t > 0. \tag{9.2-14}$$

The pdf of the increment $\Delta \triangleq X(t) - X(\tau)$ for all $t > \tau$ is given as

$$f_\Delta(\delta; t-\tau) = \frac{1}{\sqrt{2\pi\alpha(t-\tau)}} \exp\left(-\frac{\delta^2}{2\alpha(t-\tau)}\right), \tag{9.2-15}$$

since

$$E[X(t) - X(\tau)] = E[\Delta] = 0, \tag{9.2-16}$$

and

$$E\left[(X(t) - X(\tau))^2\right] = \alpha(t-\tau) \quad \text{for } t > \tau. \tag{9.2-17}$$

Example 9.2-3

(*sample functions*) We can use MATLAB to visually investigate the sample functions typical of the Wiener process. Since it is a computer simulation, we also can evaluate the effect of the limiting sequence occurring as $s = \sqrt{\alpha T}$ approaches 0 for fixed $\alpha > 0$.

We start with a 1000-element vector that is a realization of the Bernoulli random vector W with $p = 0.5$ generated as

```
u = rand(1000,1)
w = 0.5 >= u
```

The following line then converts the range of w to $\pm s$ for a prespecified value s:

```
w = s*(2*w - 1.0)
```

and then we generate a segment of a sample function of $X_T(nT) = X[n]$ as elements of the random vector

```
x = cumsum(w)
```

For the numerical experiment let $\alpha = 1.0$ and set $T = 0.01$ ($s = 0.1$). Using a computer variable x with dimension 1000 for $T = 0.01$, we get the results shown in Figure 9.2-7. Note particularly in this near limiting case, the effects of increasing variance with time. Also note that *trends* or *long-term waves* appear to develop as time progresses.

From the first-order pdf of X and the density of the increment Δ, it is possible to calculate a complete set of consistent nth-order pdf's as we have seen before. It thus follows that all nth-order pdf's of a Wiener process are Gaussian.

Definition 9.2-3 If for all positive integers n, the nth-order pdf's of a random process are all jointly Gaussian, then the process is called a *Gaussian random process.* ■

The Wiener process is thus an example of a Gaussian random process. The covariance function of the Wiener process (which is also its correlation function because $\mu_X(t) = 0$) is given as

$$K_{XX}(t_1, t_2) = \alpha \min(t_1, t_2), \quad \alpha > 0. \tag{9.2-18}$$

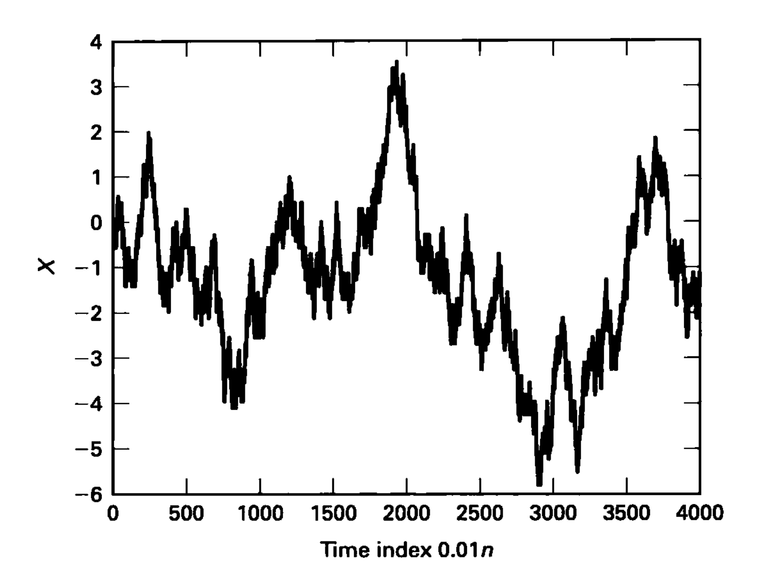


Figure 9.2-7 A Wiener process sample function approximation for $\alpha = 1$ calculated with $T = 0.01$.

To show this we take $t_1 \geq t_2$, and noting that the (forward) increment $X(t_1) - X(t_2)$ is independent of $X(t_2)$ and that they both have zero mean,

$$E[(X(t_1) - X(t_2))X(t_2)] = E[X(t_1) - X(t_2)]E[X(t_2)]$$
$$= 0$$

or

$$E[X(t_1)X(t_2)] = E[X^2(t_2)]$$
$$= \alpha t_2.$$

If $t_2 > t_1$, we get $E[X(t_2)X(t_1)] = \alpha t_1$, thus establishing Equation 9.2-18.

Note that the Wiener process has the same variance function as the Poisson process, even though the two processes are dramatically different. While the Poisson process consists solely of jumps separated by constant values, the Wiener process has no jumps and can in fact be proven to be a.s. continuous; that is, the sample functions are continuous with probability 1. Later, we will show that the Wiener process is continuous in a weaker mean-square sense (specified more precisely in Chapter 10).

Markov Random Processes

We have discussed five random processes thus far. Of these, the Wiener and Poisson are fundamental in that many other rather general random processes have been shown to be obtainable by nonlinear transformations on these two basic processes. In both cases, the difficulty of specifying a consistent set of nth-order distributions from processes with dependence was overcome by use of the independent-increments property. In fact, this is quite a

general approach in that we can start out with some arbitrary first-order distribution and then specify a distribution for the increment, thereby obtaining a consistent set of nth-order distributions that exhibit dependence.

Another way of going from the first-order probability to a consistent set of nth-order probabilities, which has proved quite useful, is the Markov process approach. Here we start with a first-order density (or PMF) and a conditional density (or conditional PMF)

$$f_X(x;t) \quad \text{and} \quad f_X(x_2|x_1;t_2,t_1), \qquad t_2 > t_1,$$

and then build up the nth-order pdf $f(x_1,\ldots,x_n;t_1,\ldots,t_n)$ (or PMF) as the product,

$$f(x_1;t)f(x_2|x_1;t_2,t_1)\ldots f(x_n|x_{n-1};t_n,t_{n-1}). \tag{9.2-19}$$

We ask the reader to show that this is a valid nth-order pdf (i.e., that this function is nonnegative and integrates to one) whenever the conditional and first-order pdf's are well defined.

Conversely, if we start with an arbitrary nth-order pdf and repeatedly use the definition of conditional probability we obtain,

$$\begin{aligned} f(x_1,\ldots,x_n;t_1,\ldots,t_n) = f(x_1;t_1)f(x_2|x_1;t_2,t_1)f(x_3|x_2,x_1;t_3,t_2,t_1) \times \\ \ldots \times f(x_n|x_{n-1},\ldots,x_1;t_n,\ldots,t_1), \end{aligned} \tag{9.2-20}$$

which can be made equivalent to Equation 9.2-19 by constraining the conditional densities to depend only on the most recent conditioning value. This motivates the following definition of a Markov random process.

Definition 9.2-4 (Markov random process)

(a) A *continuous-valued (first-order) Markov process* $X(t)$ satisfies the conditional PMF expression

$$f_X(x_n|x_{n-1},x_{n-2},\ldots,x_1;t_n,\ldots,t_1) = f_X(x_n|x_{n-1};t_n,t_{n-1}),$$

for all $x_1,x_2,\ldots,x_n$, for all $t_1 < t_2 < \ldots < t_n$, and for all integers $n > 0$.

(b) A *discrete-valued (first-order) Markov random process* satisfies the conditional PMF expression

$$P_X(x_n|x_{n-1},\ldots,x_1;t_n,\ldots,t_1) = P_X(x_n|x_{n-1};t_n,t_{n-1})$$

for all $x_1,\ldots,x_n$, for all $t_1 < \ldots < t_n$, and for all integers $n > 0$. ■

The value of the process $X(t)$ at a given time t thus determines the conditional probabilities for future values of the process. The values of the process are called the *states of the process*, and the conditional probabilities are thought of as *transition probabilities* between the states. If only a finite or countable set of values x_i is allowed, the discrete-valued Markov process is called a *Markov chain*. An example of a Markov chain is the Poisson counting process studied earlier. The Wiener process is an example of a continuous-valued Markov

process. Both these processes are Markov because of their independent-increments property. In fact, any independent-increment process is also Markov. To see this note that, for the discrete-valued case, for example,

$$
\begin{aligned}
&P_X(x_n|x_{n-1},\ldots,x_1;t_n,\ldots,t_1)\\
&\quad = P[X(t_n)=x_n|X(t_{n-1})=x_{n-1},\ldots,X(t_1)=x_1]\\
&\quad = P[X(t_n)-X(t_{n-1})=x_n-x_{n-1}|X(t_{n-1})=x_{n-1},\ldots,X(t_1)=x_1]\\
&\quad = P[X(t_n)-X(t_{n-1})=x_n-x_{n-1}] \quad \text{by the independent-increments property}\\
&\quad = P[X(t_n)-X(t_{n-1})=x_n-x_{n-1}|X(t_{n-1})=x_{n-1}] \quad \text{again by independent increments}\\
&\quad = P[X(t_n)=x_n|X(t_{n-1})=x_{n-1}]\\
&\quad = P_X(x_n|x_{n-1};t_n,t_1).
\end{aligned}
$$

Note, however, that the inverse argument is not true. A Markov random process does not necessarily have independent increments. (See Problem 9.17.)

Markov random processes find application in many areas including signal processing, communications, and control systems. Markov chains are used in communications, computer networks, and reliability theory.

Example 9.2-4

(*multiprocessor reliability*) Given a computer with two independent processors, we can model it as a three-state system: 0—both processors down; 1—exactly one processor up; and 2—both processors up. We would like to know the probabilities of these three states. A common probabilistic model is that the processors will fail randomly with time-to-failure, the *failure time*, exponentially distributed with some parameter $\lambda > 0$. Once a processor fails, the time to service it, the *service time*, will be assumed to be also exponentially distributed with parameter $\mu > 0$. Furthermore, we assume that the processor's failures and servicing are independent; thus we make the failure and service times in our probabilistic model jointly independent.

If we define $X(t)$ as the state of the system at time t, then X is a continuous-time Markov chain. We can show this by first showing that the times between state transitions of X are exponentially distributed and then invoking the memoryless property of the exponential distribution (see Problem 9.8). Analyzing the *transition times* (either failure times or service times), we proceed as follows. The transition time for going from state $X = 0$ to $X = 1$ is the minimum of two exponentially distributed service times, which are assumed to be independent. By Problem 3.26, this time will be also exponentially distributed with parameter 2μ. The expected time for this transition will thus be $1/(2\mu) = \frac{1}{2}(1/\mu)$, that is, one-half the average time to service a single processor. This is quite reasonable since both processors are down in state $X = 0$ and hence both are being serviced independently and simultaneously. The rate parameter for the transition 0 to 1 is thus 2μ. The transition 1 to 2 awaits one exponential service time at rate μ. Thus its rate is also μ. Similarly, the state transition 1 to 0 awaits only one failure at rate λ, while the transition 2 to 1 awaits the minimum of two exponentially distributed failure times. Thus its rate

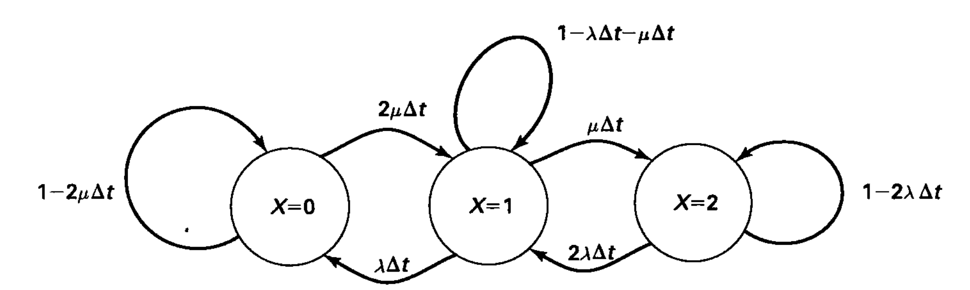


Figure 9.2-8 Short-time state-transition diagram with indicated transition probabilities.

is 2λ. Simultaneous transitions from 0 to 2 and 2 to 0 are of probability 0 and hence are ignored.

This Markov chain model is summarized in the short-time *state-transition diagram* of Figure 9.2-8. In this diagram the directed branches represent short-time, that is, as $\Delta t \to 0$, transition probabilities between the states. The transition times are assumed to be exponentially distributed with the parameter given by the branch label. These transition times might be more properly called intertransition times and are analogous to the interarrival times of the Poisson counting process, which are also exponentially distributed.

Consider the probability of being in *state* 2 at $t + \Delta t$, having been in *state* 1 at time t. This requires that the service time T_s lies in the interval $(t, t+\Delta t]$ conditional on $T_s \geq t$. Let $P_i(t) \triangleq P[X(t) = i]$ for $0 \leq i \leq 2$. Then

$$P_2(t+\Delta t) = P_1(t)P[t < T_s \leq t + \Delta t | T_s \geq t],$$

where

$$P[t < T_s \leq t + \Delta t | T_s \geq t] = \frac{F_{T_s}(t+\Delta t) - F_{T_s}(t)}{1 - F_{T_s}(t)} = \mu \Delta t + o(\Delta t).$$

Using this type of argument for connecting the probability of transitions from states at time t to states at time $t + \Delta t$ and ignoring transitions from state 2 to state 0 and vice versa enables us to write the state probability at time $t + \Delta t$ in terms of the state probability at t in vector matrix form:

$$\begin{bmatrix} P_0(t+\Delta t) \\ P_1(t+\Delta t) \\ P_2(t+\Delta t) \end{bmatrix} = \begin{bmatrix} 1-2\mu\Delta t & \lambda\Delta t & 0 \\ 2\mu\Delta t & 1-(\lambda+\mu)\Delta t & 2\lambda\Delta t \\ 0 & \mu\Delta t & 1-2\lambda\Delta t \end{bmatrix} \begin{bmatrix} P_0(t) \\ P_1(t) \\ P_2(t) \end{bmatrix} + \mathbf{o}(\Delta t),$$

where $\mathbf{o}(\Delta t)$ denotes a quantity of lower order than Δt.

Rearranging, we have

$$\begin{bmatrix} P_0(t+\Delta t) - P_0(t) \\ P_1(t+\Delta t) - P_1(t) \\ P_2(t+\Delta t) - P_2(t) \end{bmatrix} = \begin{bmatrix} -2\mu & \lambda & 0 \\ 2\mu & -(\lambda+\mu) & 2\lambda \\ 0 & \mu & -2\lambda \end{bmatrix} \begin{bmatrix} P_0(t) \\ P_1(t) \\ P_2(t) \end{bmatrix} \Delta t + \mathbf{o}(\Delta t).$$

Dividing both sides by Δt and using an obvious matrix notation, we obtain

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{A}\mathbf{P}(t). \tag{9.2-21}$$

The matrix $\mathbf{A}$ is called the *generator* of the Markov chain X. This first-order vector differential equation can be solved for an initial probability vector, $\mathbf{P}(0) \triangleq \mathbf{P}_0$, using methods of linear-system theory [9-3]. The solution is expressed in terms of the matrix exponential

$$e^{\mathbf{A}t} \triangleq \mathbf{I} + \mathbf{A}t + \frac{1}{2!}(\mathbf{A}t)^2 + \frac{1}{3!}(\mathbf{A}t)^3 + \ldots,$$

which converges for all finite t. The solution $\mathbf{P}(t)$ is then given as

$$\mathbf{P}(t) = e^{\mathbf{A}t}\mathbf{P}_0, \qquad t \geq 0.$$

For details on this method as well as how to obtain an explicit solution, see [9-4].

For the present we content ourselves with the steady-state solution obtained by setting the time derivative in Equation 9.2-21 to zero, thus yielding $\mathbf{AP}=\mathbf{0}$. From the first and last rows we get

$$-2\mu P_0 + \lambda P_1 = 0$$

and

$$+\mu P_1 - 2\lambda P_2 = 0.$$

From this we obtain $P_1 = (2\mu/\lambda)P_0$ and $P_2 = (\mu/2\lambda)P_1 = (\mu/\lambda)^2 P_0$. Then invoking $P_0 + P_1 + P_2 = 1$, we obtain $P_0 = \lambda^2/(\lambda^2 + 2\mu\lambda + \mu^2)$ and finally

$$\mathbf{P} = \frac{1}{\lambda^2 + 2\mu\lambda + \mu^2}[\lambda^2, 2\mu\lambda, \mu^2]^T.$$

Thus the steady-state probability of both processors being down is $P_0 = [\lambda/(\lambda + \mu)]^2$. Incidentally, if we had used only one processor modeled by a two-state Markov chain, we would have obtained $P_0 = \lambda/(\lambda + \mu)$.

Clearly we can generalize this example to any number of states n with independent exponential interarrival times between these states. In fact, such a process is called a *queueing process.* Other examples are the number of toll booths busy on a superhighway and congestion states in a computer or telephone network. For more on queueing systems, see [9-2]. An important point to notice in the last example is that the exponential transition times were crucial in showing the Markov property. In fact, any other distribution but exponential would not be memoryless, and the resulting state-transition process would not be a Markov chain.

Birth–Death Markov Chains

A Markov chain in which transitions are permissible only between adjacent states is called a birth–death chain. We first deal with the case where the number of states is infinite and afterwards treat the finite-state case.

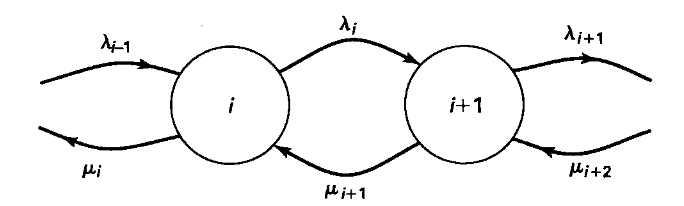


Figure 9.2-9 Markov state diagram for the birth–death process showing transition rate parameters.

1. Infinite-length queues. The state-transition diagram for the infinite-length queue is shown in Figure 9.2-9.† In going from state i to state $i+1$, we say that a *birth* has occurred. Likewise, in going from state i to state $i-1$ we say a *death* has occurred. At any time t, $P_j(t)$ is the probability of being in state j, that is of having a "population" of size j, in other words the excess of the number of births over deaths. In this model, births are generated by a Poisson process. The times between births τ_B, and the time between deaths τ_D, depend on the states but obey the exponential distribution with parameters λ_i and μ_i, respectively. The model is used widely in queuing theory where a birth is an arrival to the queue and a death is a departure of one from the queue because of the completion of service. An example is people waiting in line to purchase a ticket at a single-server ticket booth. If the theater is very large and there are no restrictions on the length of the queue (e.g., the queue may block the sidewalk and create a hazard), overflow and saturation can be disregarded. Then the dynamics of the queue are described by the basic equation $W_n = \max\{0, W_{n-1} + \tau_s - \tau_i\}$, where W_n is the waiting time in the queue for the nth arrival, τ_s is the service time for the $(n-1)$st arrival, and τ_i is the interarrival time between the nth and $(n-1)$st arrivals. This is an example of unrestricted queue length. On the other hand data packets stored in a finite-size buffer memory present a different problem. When the buffer is filled (saturation), a new arrival must be turned away (in this case we say the datum packet is "lost").

Following the procedure in Example 9.2-4, we can write that

$$\mathbf{P}(t+\Delta t) = \mathbf{B}\mathbf{P}(t),$$

where

$$\mathbf{B} = \begin{bmatrix} 1-\lambda_0\Delta t & \mu_1\Delta t & 0 & \cdots & \\ \lambda_0\Delta t & 1-(\lambda_1+\mu_1)\Delta t & \mu_2\Delta t & 0 & \cdots \\ 0 & \lambda_1\Delta t & 1-(\lambda_2+\mu_2)\Delta t & \mu_3\Delta t & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Rearranging and dividing by Δt and letting $\Delta t \to 0$, we get

$$d\mathbf{P}(t)/dt = \mathbf{A}\mathbf{P}(t),$$

†In keeping with standard practice, we draw the diagram *showing only the transition rate parameters* that is, the μ_i's and λ_i's over the links between states. This type of diagram does not show explicitly, for example, that in the Poisson case the short-time probability of staying in state i is $1-(\lambda_i+\mu_i)\Delta t$. While this type of diagram is less clear, it is less crowded than, say, the nonstandard short-time transition probability diagram in Figure 9.2-8.

where $\mathbf{P}(t) = [P_0(t), P_1(t), \ldots, P_j(t), \ldots]^T$, and $\mathbf{A}$, the *generator matrix* for the Markov chain is given by

$$\mathbf{A} = \begin{bmatrix} -\lambda_0 & \mu_1 & 0 & \cdots & \\ \lambda_0 & -(\lambda_1+\mu_1) & \mu_2 & 0 & \cdots \\ 0 & \lambda_1 & -(\lambda_2+\mu_2) & \mu_2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

In the steady state $\mathbf{P}'(t) = \mathbf{0}$. Thus, we obtain from $\mathbf{AP} = \mathbf{0}$,

$$P_1 = \rho_1 P_0,$$

$$P_2 = \rho_2 P_1 = \rho_1 \rho_2 P_0,$$

$$\vdots$$

$$P_j = \rho_j P_{j-1} = \rho_j \cdots \rho_2 \rho_1 P_0,$$

where $\rho_j \triangleq \lambda_{j-1}/\mu_j$, for $j \geq 1$.

Assuming that the series converges, we require that $\sum_{i=0}^{\infty} P_i = 1$. With the notation $r_j \triangleq \rho_j \cdots \rho_2 \rho_1$, and $r_0 = 1$, this means $P_0 \sum_{i=0}^{\infty} r_i = 1$ or $P_0 = 1/\sum_{i=0}^{\infty} r_i$. Hence the steady-state probabilities for the birth–death Markov chain are given by

$$P_j = r_j \Big/ \sum_{i=0}^{\infty} r_i, \quad j \geq 0.$$

Failure of the denominator to converge implies that there is no steady state and therefore the steady-state probabilities are zero. This model is often called the *M/M/1 queue.*

2. $M/M/1$ Queue with constant birth and death parameters and finite storage L. Here we assume that $\lambda_i = \lambda$ and $\mu_i = \mu$, for all i, and that the queue length cannot exceed L. This stochastic model can apply to the analysis of a finite buffer as shown in Figure 9.2-10. The dynamical equations are

$$dP_0(t)/dt = -\lambda P_0(t) + \mu P_1(t)$$

$$dP_1(t)/dt = +\lambda P_0(t) - (\lambda + \mu) P_1(t) + \mu P_2(t)$$

$$\vdots \quad \vdots \quad \vdots$$

$$dP_L(t)/dt = +\lambda P_{L-1}(t) - \mu P_L(t).$$

Note that the first and last equations contain only two terms, since a death cannot occur in an empty queue and a birth cannot occur when the queue has its maximum size L. From these equations, we easily obtain that the steady-state solution is $P_i = \rho^i P_0$, for $0 \leq i \leq L$, where $\rho \triangleq \lambda/\mu$. From the condition that the buffer must be in some state, we obtain that

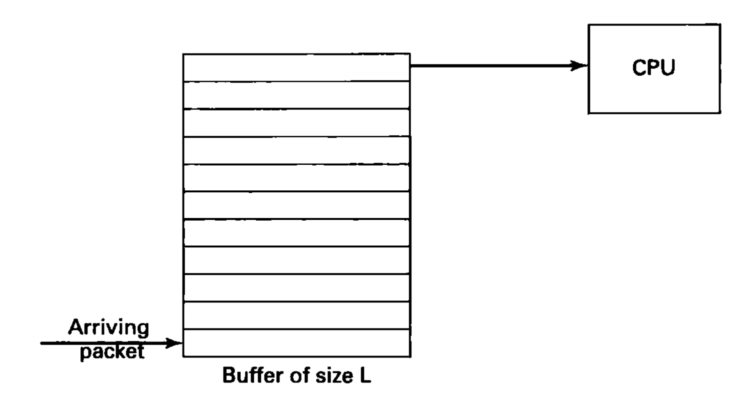


Figure 9.2-10 Illustration of packet arriving at buffer of finite size L.

$\sum_{i=0}^{L} \rho^i P_0 = 1$, or that $P_0 = (1-\rho)/(1-\rho^{L+1})$. *Saturation* occurs when the buffer is full. The steady-state probability of this event is $P_L = \rho^L(1-\rho)/(1-\rho^{L+1})$. Thus for a birth rate which is half the death rate, and a buffer of size of 10, the probability of saturation is, approximately, 5×10^{-4}.

Example 9.2-5

(*average queue size*) In computer and communication networks, *packet switching* refers to the transmission of blocks of data called packets from node to node. At each node the packets are processed with a view toward determining the next link in the source-to-destination route. The arrival time of the packets, the amount of time they have to wait in a buffer, and the service time in the CPU (the central processing unit) are random variables.

Assume a first-come, first served, infinite-capacity buffer, with exponential service time with parameter μ, and Poisson-distributed arrivals with a Poisson rate parameter of λ arrivals per unit time. We know from earlier in this section that the interarrival times of the Poisson process are i.i.d. exponential random variables with parameter λ. The state diagram for this case is identical to that of Figure 9.2-9 except that $\mu_1 = \mu_2 = \ldots = \mu$ and $\lambda_0 = \lambda_1 = \ldots = \lambda$. Then specializing the results of the previous discussion to this example, we find that $P_i = \rho^i P_0$, for $0 \leq i$, where $\rho \triangleq \lambda/\mu$ and $P_0 = (1-\rho)$. Thus, in the steady state $P_i = \rho^i(1-\rho)$, and the average number of packets in the queue $E[N]$, is computed from

$$E[N] = \sum_{i=0}^{\infty} iP_i = \frac{\rho}{(1-\rho)^2}.$$

We leave the details of this elementary calculation as an exercise.

Example 9.2-6

(*finite capacity buffer*) We revisit Example 9.2-5 except that now the arriving data packets are stored in a buffer of size L. Consider the following set-up: The data stored in the buffer are processed by a CPU on a first-come, first-service basis.

Assume that, say at time t, the buffer is filled to capacity, and that there is a packet being processed in the CPU and an arriving packet on the way to the buffer. If the interarrival time τ_i between this packet and the previous one is less than τ_s, the service time for the packet in the CPU, the arriving packet will be lost. The probability of this event is

$$\begin{aligned} P[\text{"packet loss"}] &= P\left[\text{"saturation"} \cap \{\tau_s > \tau_i\}\right] \\ &= \rho^L(1-\rho)/(1-\rho^{L+1}) \times P[\tau_s - \tau_i > 0], \end{aligned}$$

since the event's "saturation" and $\{\tau_s > \tau_i\}$ are independent. Since τ_s and τ_i are independent, the probability $P[\tau_s - \tau_i > 0]$ can easily be computed by convolution. The result is $P[\tau_s - \tau_i > 0] = \lambda/(\lambda + \mu)$. The probability of losing the incoming packet is then

$$P[\text{"packet loss"}] = \rho^L(1-\rho)/(1-\rho^{L+1}) \times \rho/(1+\rho),$$

which, for $\rho = 0.5$, yields $P[\text{"packet loss"}] = 1.6 \times 10^{-4}$ for the buffer of size 10, with arrival rate equal to half the service rate.

Chapman–Kolmogorov Equations

In the examples of a Markov random sequence in Chapter 8, we specified the transition density as a one-step transition, that is, from $n-1$ to n. More generally, we can specify the transition density from time n to time $n+k$, where $k \geq 0$, as in the general definition of a Markov random sequence. However, in this more general case we must make sure that this multistep transition density is consistent, that is, that there exists a one-step density that would sequentially yield the same results. This problem is even more important in the random process case, where due to continuous time one is always effectively considering multistep transition densities; that is, between any two times $t_2 \neq t_1$, there is a time in between.

For example, given a continuous-time transition density $f_X(x_2|x_1; t_2, t_1)$, how do we know that an unconditional pdf $f_X(x;t)$ can be found to satisfy the equation

$$f_X(x_2;t_2) = \int_{-\infty}^{+\infty} f_X(x_2|x_1;t_2,t_1) f_X(x_1;t_1)dx_1$$

for all $t_2 > t_1$, and all x_1 and x_2?

The Chapman–Kolmogorov equations supply both necessary and sufficient conditions for these general transition densities. There is also a version of the Chapman–Kolmogorov equations for the discrete-valued case involving PMFs of multistep transitions.

Consider three times $t_3 > t_2 > t_1$ and the Markov process random variables at these three times $X(t_3)$, $X(t_2)$, and $X(t_1)$. We wish to compute the conditional density of $X(t_3)$ given $X(t_1)$. First, we write the joint pdf

$$f_X(x_3, x_1; t_3, t_1) = \int_{-\infty}^{+\infty} f_X(x_3|x_2, x_1; t_3, t_2, t_1) f_X(x_2, x_1; t_2, t_1)dx_2.$$

If we now divide both sides of this equation by $f(x_1;t_1)$, we obtain

$$f_X(x_3|x_1) = \int_{-\infty}^{+\infty} f_X(x_3|x_2, x_1) f_X(x_2|x_1) dx_2,$$

where we have suppressed the times t_i for notational simplicity. Then using the Markov property the above becomes

$$f_X(x_3|x_1) = \int_{-\infty}^{+\infty} f_X(x_3|x_2) f_X(x_2|x_1) dx_2, \tag{9.2-22}$$

which is known as the *Chapman–Kolmogorov equation* for the transition density $f_X(x_3|x_1)$ of a Markov process. This equation must hold for all $t_3 > t_2 > t_1$ and for all values of x_3 and x_1. It can be proven that the Chapman–Kolmogorov condition expressed in Equation 9.2-22 is also sufficient for the existence of the transition density in question [9-5].

Random Process Generated from Random Sequences

We can obtain a Markov random process as the limit of an infinite number of simulations of Markov random sequences. For example, consider the random sequence generated by the equation

$$X[n] = \rho X[n-1] + W[n], \quad -\infty < n < +\infty,$$

as given in Example 8.4-6 of Chapter 8, where $|\rho| < 1.0$ to ensure stability. There we found that the correlation function of $X[n]$ was

$$R_{XX}[m] = \sigma_W^2 \, \rho^{|m|},$$

where σ_W^2 is the variance of the independent random sequence $W[n]$. Replacing $X[n]$ with $X(nT)$, and setting $X(t) = X[nT]$ for $nT \leq t < (n+1)T$, we get

$$R_{XX}(t+\tau, t) = \sigma_W^2 \; \rho^{|\tau/T|} = \sigma_W^2 \; \exp(-\alpha|\tau|),$$

where $\alpha \triangleq \frac{1}{T} \ln \frac{1}{\rho}$ or alternatively $\rho = \exp(-\alpha T)$. Thus, if we generate a set of simulations with $T_k \triangleq T_0/k$ for $k = 1, 2, 3, \ldots$, and then for each simulation set $\rho_k \triangleq \sqrt[k]{\exp(-\alpha T_0)}$, we will get a set of denser and denser approximations to a limiting random process $X(t)$, that is WSS with correlation function

$$R_{XX}(t+\tau, t) = \sigma_W^2 \; \exp(-\alpha|\tau|).$$

9.3 CONTINUOUS-TIME LINEAR SYSTEMS WITH RANDOM INPUTS

In this section we look at transformations of stochastic processes. We concentrate on the case of linear transformations with memory, since the memoryless case can be handled by

the transformation of random variables method of Chapter 3. The definition of a linear continuous-time system is recalled first.

Definition 9.3-1 Let $x_1(t)$ and $x_2(t)$ be two deterministic time functions and let a_1 and a_2 be two scalar constants. Let the linear system be described by the operator equation $y = L\{x\}$. Then the system is linear if

$$L\{a_1x_1(t) + a_2x_2(t)\} = a_1L\{x_1(t)\} + a_2L\{x_2(t)\} \tag{9.3-1}$$

for all admissible functions x_1 and x_2 and all scalars a_1 and a_2. ■

This amounts to saying that the response to a weighted sum of inputs must be the weighted sum of the responses to each one individually. Also, in this definition we note that the inputs must be in the allowable input space for the system (operator) L. When we think of generalizing L to allow a random process input, the most natural choice is to input the sample functions of X and find the corresponding sample functions of the output, which thereby define a new random process Y. Just as the original random process X is a mapping from the sample space to a function space, the linear system in turn maps this function space to a new function space. The cascade or composition of the two maps thus defines an output random process. This is depicted graphically in Figure 9.3-1. Our goal in this section will be to find out how the first- and second-order moments, that is, the mean and correlation (and covariance), are transformed by a linear system.

Theorem 9.3-1 Let the random process $X(t)$ be the input to a linear system L with output process $Y(t)$. Then the mean function of the output is given as

$$\begin{aligned} E[Y(t)] &= L\{E[X(t)]\} \\ &= L\{\mu_X(t)\}. \end{aligned} \tag{9.3-2}$$

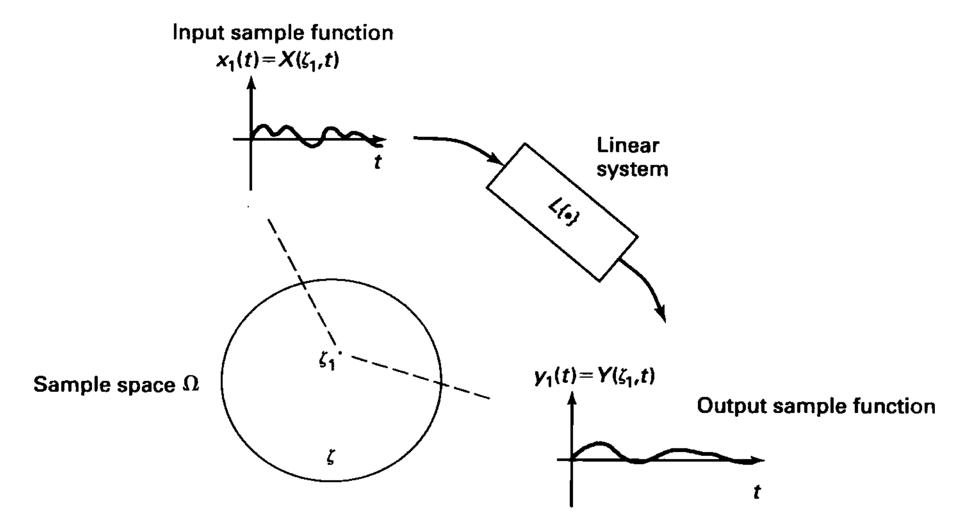


Figure 9.3-1 Interpretation of applying a random process to a linear system.

Proof (formal). By definition we have for each sample function

$$Y(t,\zeta) = L\{X(t,\zeta)\}$$

so

$$E[Y(t)] = E[L\{X(t)\}].$$

If we can interchange the two operators, we get the result that the mean function of the output is just the result of L operating on the mean function of the input. This can be heuristically (formally) justified as follows, if we assume the operator L can be represented by the superposition integral:

$$Y(t) = \int_{-\infty}^{+\infty} h(t,\tau)X(\tau)d\tau.$$

Taking the expectation, we obtain

$$\begin{aligned} E[Y(t)] &= E\left[\int_{-\infty}^{+\infty} h(t,\tau)X(\tau)d\tau\right] \\ &= \int_{-\infty}^{+\infty} h(t,\tau)E[X(\tau)]d\tau \\ &= L\{\mu_X(t)\}. \quad \blacksquare \end{aligned}$$

We present a rigorous proof of this theorem after we study the mean-square stochastic integral in Chapter 10. For now, we will assume it is valid, and next look at how the correlation function is transformed by a linear system. There are now two stochastic processes to consider, the input and the output, and the cross-correlation function $E[X(t_1)Y^*(t_2)]$ comes into play. We thus define the *cross-correlation function*

$$R_{XY}(t_1,t_2) \triangleq E[X(t_1)Y^*(t_2)].$$

From the autocorrelation function of the input $R_{XX}(t_1,t_2)$, we first calculate the cross-correlation function $R_{XY}(t_1,t_2)$ and then the autocorrelation function of the output $R_{YY}(t_1,t_2)$. If the mean is zero for the input process, then by Theorem 9.3-1 the mean of the output process is also zero. Thus the following results can be seen also to hold for covariance functions by changing the input to the centered process $X_c(t) \triangleq X(t) - \mu_X(t)$, which produces the centered output $Y_c(t) \triangleq Y(t) - \mu_Y(t)$.

Theorem 9.3-2 Let $X(t)$ and $Y(t)$ be the input and output random processes of the linear operator L. Then the following hold:

$$R_{XY}(t_1,t_2) = L_2^*\{R_{XX}(t_1,t_2)\}, \tag{9.3-3}$$

$$R_{YY}(t_1,t_2) = L_1\{R_{XY}(t_1,t_2)\}, \tag{9.3-4}$$

where L_i means the time variable of the operator L is t_i.

Proof (formal). Write

$$\begin{aligned} X(t_1)Y^*(t_2) &= X(t_1)L_2^*\{X^*(t_2)\} \\ &= L_2^*\{X(t_1)X^*(t_2)\}, \end{aligned}$$

where we have used the *adjoint operator* L^* whose impulse response is $h^*(t,\tau)$, that is, the complex conjugate of $h(t,\tau)$. Then

$$\begin{aligned} E[X(t_1)Y^*(t_2)] &= E\left[L_2^*\{X(t_1)X^*(t_2)\}\right] \\ &= L_2^*\{E[X(t_1)X^*(t_2)]\} \quad \text{by interchanging } L_2^* \text{ and } E, \\ &= L_2^*\{R_{XX}(t_1,t_2)\}, \end{aligned}$$

which is Equation 9.3-3. Similarly, to prove Equation 9.3-4, we multiply by $Y^*(t_2)$ and get

$$Y(t_1)Y^*(t_2) = L_1\{X(t_1)Y^*(t_2)\}$$

so that

$$\begin{aligned} E[Y(t_1)Y^*(t_2)] &= E\left[L_1\{X(t_1)Y^*(t_2)\}\right] \\ &= L_1\{E[X(t_1)Y^*(t_2)]\} \quad \text{by interchanging } L_1 \text{ and } E, \\ &= L_1\{R_{XY}(t_1,t_2)\}, \end{aligned}$$

which is Equation 9.3-4. If we combine Equation 9.3-3 and Equation 9.3-4, we get

$$R_{YY}(t_1,t_2) = L_1L_2^*\{R_{XX}(t_1,t_2)\}. \quad \blacksquare \tag{9.3-5}$$

Example 9.3-1

(*edge or "change" detector*) Let $X(t)$ be a real-valued random process, modeling a certain sensor signal, and define $Y(t) \triangleq L\{X(t)\} \triangleq X(t) - X(t-1)$ so

$$E[Y(t)] = L\{\mu_X(t)\} = \mu_X(t) - \mu_X(t-1).$$

Also

$$R_{XY}(t_1,t_2) = L_2\{R_{XX}(t_1,t_2)\} = R_{XX}(t_1,t_2) - R_{XX}(t_1,t_2-1)$$

and

$$\begin{aligned} R_{YY}(t_1,t_2) = L_1\{R_{XY}(t_1,t_2)\} &= R_{XY}(t_1,t_2) - R_{XY}(t_1-1,t_2) \\ &= R_{XX}(t_1,t_2) - R_{XX}(t_1-1,t_2) - R_{XX}(t_1,t_2-1) \\ &\quad + R_{XX}(t_1-1,t_2-1). \end{aligned}$$

To be specific, if we take $\mu_X(t) = 0$ and

$$R_{XX}(t_1,t_2) \triangleq \sigma_X^2 \exp(-\alpha|t_1-t_2|),$$

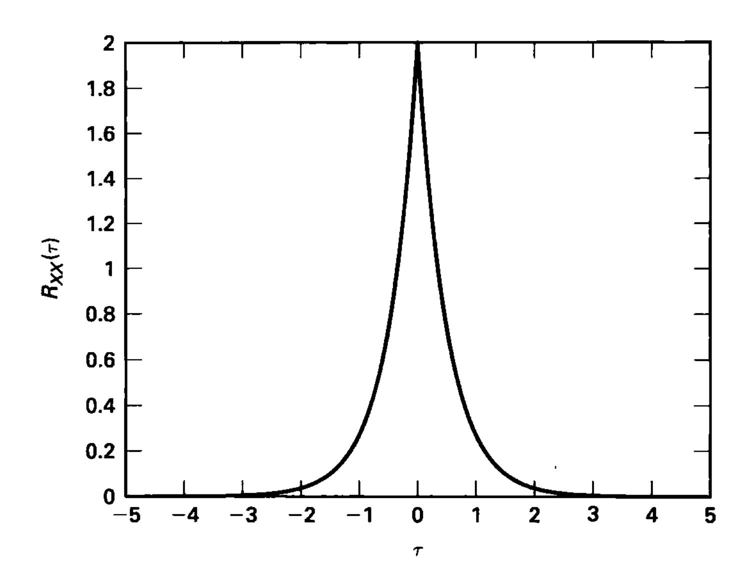


Figure 9.3-2 Input correlation function R_{XX} of Example 9.3-1 versus $\tau = t_1 - t_2$.

then

$$E[Y(t)] = 0 \qquad \text{since } \mu_X = 0,$$

and

$$R_{YY}(t_1,t_2) = \sigma_X^2\big(2\exp(-\alpha|t_1 - t_2|) - \exp(-\alpha|t_1 - t_2 - 1|) - \exp(-\alpha|t_1 - t_2 + 1|)\big).$$

We note that both R_{XX} and R_{XY} are functions only of the difference of the two observation times t_1 and t_2. The input correlation function R_{XX} is plotted in Figure 9.3-2, for $\alpha = 2$ and $\sigma_X^2 = 2$. Note the negative correlation values in output correlation function R_{YY}, shown in Figure 9.3-3, introduced by the difference operation of the edge detector. The variance of $Y(t)$ is constant and is given as

$$\sigma_Y^2(t) = \sigma_Y^2 = 2\sigma_X^2[1 - \exp(-\alpha)].$$

We see that as α tends to zero, the variance of Y goes to zero. This is because as α tends to zero, $X(t)$ and $X(t-1)$ become very positively correlated, and hence there is very little power in their difference.

Example 9.3-2

(*derivative process*) Let $X(t)$ be a real-valued random process with constant mean function $\mu_X(t) = \mu$ and covariance function

$$K_{XX}(t,s) = \sigma^2 \cos\omega_0(t-s).$$

We wish to determine the mean and covariance function of the derivative process $X'(t)$. Here the linear operator is $d(\cdot)/dt$. First we determine the mean,

$$\mu_{X'}(t) = E[X'(t)] = \frac{d}{dt}E[X(t)] = \frac{d}{dt}\mu_X(t) = \frac{d}{dt}\mu = 0.$$

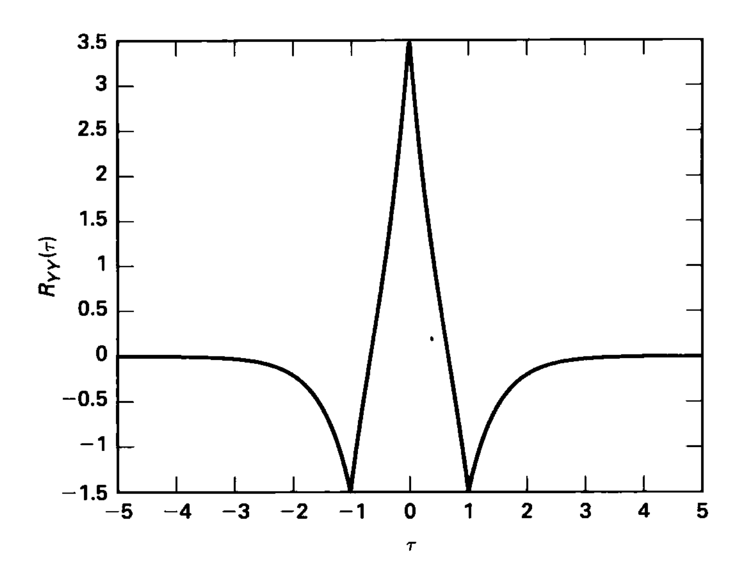


Figure 9.3-3 Output correlation function R_{YY} of Example 9.3-1 versus $\tau = t_1 - t_2$.

Now, for this real-valued random process, the covariance function of $X'(t)$ is

$$K_{X'X'}(t_1, t_2) = E[X'(t_1)X'(t_2)],$$

since $\mu'_X(t) = 0$. Thus by Equation 9.3-5, with $X'(t) = Y(t)$,

$$K_{X'X'}(t_1, t_2) = \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2} K_{XX}(t_1, t_2)\right) = \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2}\sigma^2 \cos \omega_0(t_1 - t_2)\right)$$

$$= \frac{\partial}{\partial t_1}(\omega_0 \sigma^2 \sin \omega_0(t_1 - t_2)) = (\omega_0 \sigma)^2 \cos \omega_0(t_1 - t_2).$$

We note that the result is just the original covariance function scaled up by the factor ω_0^2. This similarity in form happened because the given $K_{XX}(t, s)$ is the covariance function of a sine wave with random amplitude and phase (cf. Example 9.1-5). Since the phase is random, the sine and its derivative the cosine are indistinguishable by shape.

White Noise

Let the random process under consideration be the Wiener process of Section 9.2. Here we consider the derivative of this process. For any $\alpha > 0$, the covariance function of the Wiener process is $K_{XX}(t_1, t_2) = \alpha \min(t_1, t_2)$ and its mean function $\mu_X = 0$. Let $W(t) = dX(t)/dt$. Then proceeding as in the above example, we can calculate $\mu_W(t) = E[dX(t)/dt] = d\mu_X(t)/dt = 0$. For the covariance,

$$\begin{aligned}
K_{WW}(t_1,t_2) &= \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2}K_{XX}(t_1,t_2)\right) \\
&= \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2}\alpha\min(t_1,t_2)\right) \\
&= \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2}\begin{cases}\alpha t_2, & t_2 < t_1 \\ \alpha t_1, & t_2 \geq t_1\end{cases}\right) \\
&= \frac{\partial}{\partial t_1}\begin{cases}\alpha, & t_2 < t_1 \\ 0, & t_2 > t_1\end{cases} \\
&= \frac{\partial}{\partial t_1}\begin{cases}0, & t_1 < t_2 \\ \alpha, & t_1 > t_2\end{cases} \\
&= \frac{\partial}{\partial t_1}\alpha u(t_1 - t_2) \\
&= \alpha\delta(t_1 - t_2).
\end{aligned}$$

Thus the covariance function of white noise is the impulse function. Since white noise always has zero mean, the correlation function too is an impulse. It is common to see

$$R_{WW}(t_1,t_2) = \sigma^2\delta(t_1 - t_2) = K_{WW}(t_1,t_2), \tag{9.3-6}$$

with α replaced by σ^2, but one should note that the power in this process $E[|W(t)|^2] = \sigma^2\delta(0) = \infty$, not σ^2. In fact, σ^2 is a power density for the white noise process.

Note that the sample functions are highly discontinuous and the white noise process is not *separable*.†

9.4 SOME USEFUL CLASSIFICATIONS OF RANDOM PROCESSES

Here we look at several classes of random processes and pairs of processes. These classifications also apply to the random sequences studied earlier.

Definition 9.4-1 Let X and Y be random processes. They are

(a) *Uncorrelated* if $R_{XY}(t_1,t_2) = \mu_X(t_1)\mu_Y^*(t_2)$, for all t_1 and t_2;
(b) *Orthogonal* if $R_{XY}(t_1,t_2) = 0$ for all t_1 and t_2;
(c) *Independent* if for all positive integers n, the nth-order CDF of X and Y factors, that is,

$$\begin{aligned}
&F_{XY}(x_1,y_1,x_2,y_2,\ldots,x_n,y_n;t_1,\ldots,t_n) \\
&\quad = F_X(x_1,\ldots,x_n;t_1,\ldots,t_n)F_Y(y_1,\ldots,y_n;t_1,\ldots,t_n),
\end{aligned}$$

for all x_i, y_i and for all $t_1,\ldots,t_n$. ■

†The idea of separability (cf. Section 9.1) is to make a countable set of points on the t-axis (e.g., time-axis) determine the properties of the process. In effect it says that knowing the pdf over a countable set of points implies knowing the pdf everywhere. See [9-6].

Note that two random processes are orthogonal if they are uncorrelated and *at least one* of their mean functions is zero. Actually, the orthogonality concept is useful only when the random processes under consideration are zero-mean, in which case it becomes equivalent to the uncorrelated condition. The orthogonality concept was introduced for random vectors in Chapter 5. This concept will prove useful for estimating random processes and sequences in Chapter 11.

A random process may be uncorrelated, orthogonal, or independent of *itself* at earlier and/or later times. For example, we may have $R_{XX}(t_1, t_2) = 0$ for all $t_1 \neq t_2$, in which case we call X an *orthogonal random process.* Similarly $X(t)$ may be independent of $\{X(t_1), \ldots, X(t_n)\}$ for all $t \notin \{t_1, \ldots, t_n\}$ and for all $t_1, \ldots, t_n$ and for all $n \geq 1$. Then we say $X(t)$ is an *independent random process.* Clearly, the sample functions of such processes will be quite rough, since arbitrarily small changes in t yield complete independence.

Stationarity

We say a random process is stationary when its statistics do not change with the continuous parameter, often time. The formal definition is:

Definition 9.4-2 A random process $X(t)$ is *stationary* if it has the same nth-order CDF as $X(t+T)$, that is, the two n-dimensional functions

$$F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = F_X(x_1, \ldots, x_n; t_1 + T, \ldots, t_n + T)$$

are identically equal for all T, for all positive integers n, and for all $t_1, \ldots, t_n$. ■

When the CDF is differentiable, we can equivalently write this in terms of the pdf as

$$f_X(x_1, \ldots, x_n; t_1, \ldots, t_n) = f_X(x_1, \ldots, x_n; t_1 + T, \ldots, t_n + T),$$

and this is the form of the stationarity condition that is most often used. This definition implies that the mean of a stationary process is a constant. To prove this note that $f(x;t) = f(x; t+T)$ for all T implies $f(x;t) = f(x;0)$ by taking $T = -t$, which in turn implies that $E[X(t)] = \mu_X(t) = \mu_X(0)$, a constant.

Since the second-order density is also shift invariant, that is,

$$f(x_1, x_2; t_1, t_2) = f(x_1, x_2; t_1 + T, t_2 + T),$$

we have, on choosing $T = -t_2$, that

$$f(x_1, x_2; t_1, t_2) = f(x_1, x_2; t_1 - t_2, 0),$$

which implies $E[X(t_1)X^*(t_2)] = R_{XX}(t_1 - t_2, 0)$. In the stationary case, therefore, the notation for correlation function can be simplified to a function of just the shift $\tau \triangleq t_1 - t_2$ between the two sampling instants or parameters. Thus we can define the *one-parameter correlation function*

$$\begin{aligned} R_{XX}(\tau) &\triangleq R_{XX}(\tau, 0) \\ &= E[X(t+\tau)X^*(t)], \end{aligned} \tag{9.4-1}$$

which is functionally independent of the parameter t. Examples of this sort of correlation function were seen in Section 9.3.

A weaker form of stationarity exists which does not directly constrain the nth-order CDFs, but rather just the first- and second-order moments. This property, which is easier to check, is called wide-sense stationarity and will be quite useful in what follows.

Definition 9.4-3 A random process X is *wide-sense stationary* (WSS) if $E[X(t)] = \mu_X$, a constant, and $E[X(t+\tau)X^*(t)] = R_{XX}(\tau)$ for all $-\infty < \tau + \infty$, independent of the time parameter t. ■

Example 9.4-1

(*WSS complex exponential*) Let $X(t) \triangleq A\exp(j2\pi ft)$ with f a known real constant and A a *real-valued* random variable with mean $E[A] = 0$ and finite average power $E[A^2]$. Calculating the mean and correlation of $X(t)$, we obtain

$$E[X(t)] = E[A\exp(j2\pi ft)] = E[A]\exp(j2\pi ft) = 0,$$

and

$$E[X(t+\tau)X^*(t)] = E[A\exp(j2\pi f(t+\tau))A\exp(-j2\pi ft)] = E[A^2]\exp(j2\pi f\tau) = R_{XX}(\tau).$$

Note that $E[A] = 0$ is a necessary condition for WSS here. *Question*: Would this work with a cosine function in place of the complex exponential?

The process in Example 9.4-1, while shown to be wide-sense stationary, is clearly not stationary. Consider, for example, that $X(0)$ must be pure real while $X(1/(4f))$ must always be pure imaginary. We thus conclude that the WSS property is considerably weaker than stationarity.

We can generalize this example to have M complex sinusoids and obtain a rudimentary frequency domain representation for zero-mean WSS random processes. Consider

$$X(t) = \sum_{k=1}^{M} A_k \exp(j2\pi f_k t),$$

where the generally complex random variables A_k are uncorrelated with mean zero and variances σ_k^2. Then the resulting random process is WSS with mean zero and autocorrelation (or autocovariance) equal to

$$R_{XX}(\tau) = \sum_{k=1}^{M} \sigma_k^2 \exp(j2\pi f_k \tau). \qquad (9.4\text{-}2)$$

For such random processes $X(t)$, the set of random coefficients $\{A_k\}$ constitutes a frequency domain representation. From our experience with Fourier analysis of deterministic functions, we can expect that as M became large and as the f_k became dense, that is, the spacing between the f_k became small and they cover the frequency range of interest, most random processes would have such an approximate representation. Such is the case (cf. Section 10.6).

9.5 WIDE-SENSE STATIONARY PROCESSES AND LSI SYSTEMS

In this section we treat random processes that are jointly stationary and of *second order*, that is,

$$E[|X(t)|^2] < \infty.$$

Some important properties of the auto- and cross-correlation functions of stationary second-order processes are summarized as follows. They, of course, also hold for the respective covariance functions.

(1) $|R_{XX}(\tau)| \leq R_{XX}(0)$, which, for the real case, directly follows from $E[|X(t+\tau) - X(t)|^2] \geq 0$.

(2) $|R_{XY}(\tau)| \leq \sqrt{R_{XX}(0)R_{YY}(0)}$, which is derived using the Schwarz inequality. (cf. Section 4.3. Also called diagonal dominance.) It also proves the complex case of 1.

(3) $R_{XX}(\tau) = R^*_{XX}(-\tau)$, since $E[X(t+\tau)X^*(t)] = E[X(t)X^*(t-\tau)] = E^*[X(t-\tau)X^*(t)]$ for WSS random processes, which is called the *conjugate symmetry property*. In the special case of a real-valued process, this property becomes that of even symmetry, that is,

3a. $R_{XX}(\tau) = R_{XX}(-\tau)$.

Another important property of the autocorrelation function of a complex-valued, stationary random process is that it must be *positive semidefinite*, that is,

(4) for all $N > 0$, all $t_1 < t_2 < \ldots < t_N$ and all complex $a_1, a_2, \ldots, a_N$,

$$\sum_{k=1}^{N}\sum_{l=1}^{N} a_k a_l^* R_{XX}(t_k - t_l) \geq 0.$$

This was shown in Section 9.1 to be a necessary condition for a given function $g(t,s) = g(t-s)$ to be an autocorrelation function. We will show that this property is also a sufficient condition, so that positive semidefiniteness actually *characterizes* autocorrelation functions. In general, however, it is very difficult to check property (4) directly.

To start off, we can specialize the results of Theorems 9.3-1 and 9.3-2, which were derived for the general case, to LSI systems. Rewriting Equation 9.3-2 we have

$$\begin{aligned} E[Y(t)] &= L\{\mu_X(t)\} \\ &= \int_{-\infty}^{\infty} \mu_X(\tau)h(t-\tau)\,d\tau \\ &= \mu_X(t) * h(t). \end{aligned}$$

Using Theorem 9.3-2 and Equations 9.3-3 and 9.3-4, we get also

$$R_{XY}(t_1, t_2) = \int_{-\infty}^{+\infty} h^*(\tau_2) R_{XX}(t_1, t_2 - \tau_2) d\tau_2,$$

and

$$R_{YY}(t_1, t_2) = \int_{-\infty}^{+\infty} h(\tau_1) R_{XY}(t_1 - \tau_1, t_2) d\tau_1,$$

which can be written in convolution operator notation as

$$R_{XY}(t_1, t_2) = h^*(t_2) * R_{XX}(t_1, t_2),$$

where the convolution is along the t_2-axis, and

$$R_{YY}(t_1, t_2) = h(t_1) * R_{XY}(t_1, t_2),$$

where the convolution is along the t_1-axis. Combining these two equations, we get $R_{YY}(t_1, t_2) = h(t_1) * R_{XX}(t_1, t_2) * h^*(t_2)$.

Wide-Sense Stationary Case

If we input the stationary random process $X(t)$ to an LSI system with impulse response $h(t)$, then the output random process can be expressed as the convolution integral,

$$Y(t) = \int_{-\infty}^{+\infty} h(\tau) X(t - \tau) d\tau, \tag{9.5-1}$$

when this integral exists. Computing the mean of the output process $Y(t)$, we get

$$\begin{aligned} E[Y(t)] &= \int_{-\infty}^{+\infty} h(\tau) E[X(t - \tau)] d\tau \quad \text{by Theorem 9.3-1,} \\ &= \int_{-\infty}^{+\infty} h(\tau) \mu_X d\tau = \mu_X \int_{-\infty}^{+\infty} h(\tau) d\tau, \qquad (9.5\text{-}2) \\ &= \mu_X H(0), \end{aligned}$$

where $H(\omega)$ is the system's frequency response.

We thus see that the mean of the output is constant and equals the mean of the input times the system function evaluated at $\omega = 0$, the so-called "dc gain" of the system. If we compute the cross-correlation function between the input process and the output process, we find that

$$\begin{aligned} R_{YX}(\tau) &= E[Y(t + \tau) X^*(t)] \\ &= E[Y(t) X^*(t - \tau)] \quad \text{by substituting } t - \tau \text{ for } t, \\ &= \int_{-\infty}^{+\infty} h(\alpha) E[X(t - \alpha) X^*(t - \tau)] d\alpha, \end{aligned}$$

and bringing the operator E inside the integral by Theorem 9.3-2,

$$= \int_{-\infty}^{+\infty} h(\alpha) R_{XX}(\tau - \alpha) d\alpha,$$

which can be rewritten as

$$R_{YX}(\tau) = h(\tau) * R_{XX}(\tau). \tag{9.5-3}$$

Thus, the cross-correlation R_{YX} equals h convolved with the autocorrelation R_{XX}. This fact can be used to identify unknown systems (see Problem 9.28).

The output autocorrelation function $R_{YY}(\tau)$ can now be obtained from $R_{YX}(\tau)$ as follows:

$$\begin{aligned}
R_{YY}(\tau) &= E[Y(t+\tau)Y^*(t)] \\
&= E[Y(t)Y^*(t-\tau)] \quad \text{by substituting } t \text{ for } t-\tau, \\
&= \int_{-\infty}^{+\infty} h^*(\alpha)E[Y(t)X^*(t-\tau-\alpha)]d\alpha \\
&= \int_{-\infty}^{+\infty} h^*(\alpha)E[Y(t)X^*(t-(\tau+\alpha))]d\alpha \\
&= \int_{-\infty}^{+\infty} h^*(\alpha)R_{YX}(\tau+\alpha)d\alpha \\
&= \int_{-\infty}^{+\infty} h^*(-\alpha)R_{YX}(\tau-\alpha)d\alpha \\
&= h^*(-\tau) * R_{YX}(\tau).
\end{aligned}$$

Combining both equations, we get

$$R_{YY}(\tau) = h(\tau) * h^*(-\tau) * R_{XX}(\tau). \tag{9.5-4}$$

We observe that when $R_{XX}(\tau) = \delta(\tau)$, then the output correlation function is $R_{YY}(\tau) = h(\tau) * h^*(-\tau)$, which is sometimes called the *autocorrelation impulse response* (AIR) denoted as $g(\tau) = h(\tau) * h^*(-\tau)$. Note that $g(\tau)$ must be positive semidefinite, and indeed $FT\{g(\tau)\} = |H(\omega)|^2 \geq 0$.

Similarly, we also find (proof left as an exercise for the reader)

$$\begin{aligned}
R_{XY}(\tau) &= \int_{-\infty}^{+\infty} h^*(-\alpha)R_{XX}(\tau-\alpha)d\alpha \\
&= h^*(-\tau) * R_{XX}(\tau),
\end{aligned} \tag{9.5-5a}$$

and

$$\begin{aligned}
R_{YY}(\tau) &= \int_{-\infty}^{+\infty} h(\alpha)R_{XY}(\tau-\alpha)d\alpha \\
&= h(\tau) * R_{XY}(\tau) \\
&= h(\tau) * h^*(-\tau) * R_{XX}(\tau) \\
&= g(\tau) * R_{XX}(\tau).
\end{aligned}$$

This elegant and concise notation is shorthand for

$$R_{YY}(\tau) = \int_{-\infty}^{\infty} g(\tau')R_{XX}(\tau - \tau')d\tau' \quad \text{(a convolution)} \tag{9.5-5b}$$

$$g(\tau') = \int_{-\infty}^{\infty} h^*(\alpha)h(\alpha + \tau')d\alpha. \quad \text{(a correlation product)} \tag{9.5-5c}$$

Example 9.5-1

(*derivative of WSS process*) Let the second-order random process $X(t)$ be stationary with one-parameter correlation function $R_X(\tau)$ and constant mean function $\mu_X(t) = \mu_X$. Consider the system consisting of a derivative operator, that is,

$$Y(t) = \frac{dX(t)}{dt}.$$

Using the above equations, we find $\mu_Y(t) = d\mu_X(t)/dt = 0$ and cross-correlation function

$$\begin{aligned} R_{XY}(\tau) &= u_1^*(-\tau) * R_{XX}(\tau) \\ &= -\frac{dR_{XX}(\tau)}{d\tau}, \end{aligned}$$

since the impulse response of the derivative operator is $h(t) = d\delta(t)/dt = u_1(t)$, the (formal) derivative of the impulse $\delta(t)$, sometimes called the unit doublet.†

$$\begin{aligned} R_{YY}(\tau) &= u_1(\tau) * R_{XY}(\tau) \\ &= \frac{dR_{XY}(\tau)}{d\tau} \\ &= -\frac{d^2R_{XX}(\tau)}{d\tau^2}. \end{aligned}$$

Notice the AIR function here is $g(\tau) = -u_2(\tau)$, minus the second (formal) derivative of $\delta(\tau)$.

Power Spectral Density

For WSS, and hence for stationary processes, we can define a useful density for average power versus frequency, called the *power spectral density* (psd).

Definition 9.5-1 Let $R_{XX}(\tau)$ be an autocorrelation function. Then we define the *power spectral density* $S_{XX}(\omega)$ to be its Fourier transform (if it exists), that is,

$$S_{XX}(\omega) \triangleq \int_{-\infty}^{+\infty} R_{XX}(\tau)e^{-j\omega\tau}\,d\tau. \quad \blacksquare \tag{9.5-6}$$

Under quite general conditions one can define the inverse Fourier transform, which equals $R_{XX}(\tau)$ at all points of continuity,

$$R_{XX}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{XX}(\omega)e^{+j\omega\tau}\,d\omega. \tag{9.5-7}$$

†In this u-function notation, $u_{-1}(t) = u(t)$ the unit step function, and $u_0(t) = \delta(t)$ the unit impulse [9-9].

Table 9.5-1 Correlation Function Properties of Corresponding Power Spectral Densities

Random Process	Correlation Function	Power Spectral Density
$X(t)$	$R_{XX}(\tau)$	$S_{XX}(\omega)$
$aX(t)$	$\|a\|^2 R_{XX}(\tau)$	$\|a\|^2 S_{XX}(\omega)$
$X_1(t)+X_2(t)$ with X_1 and X_2 orthogonal	$R_{X_1X_1}(\tau)+R_{X_2X_2}(\tau)$	$S_{X_1X_1}(\omega)+S_{X_2X_2}(\omega)$
$X'(t)$	$-d^2R_{XX}(\tau)/d\tau^2$	$\omega^2 S_{XX}(\omega)$
$X^{(n)}(t)$	$(-1)^n d^{2n}R_{XX}(\tau)/d\tau^{2n}$	$\omega^{2n}S_{XX}(\omega)$
$X(t)\exp(j\omega_0 t)$	$\exp(j\omega_0\tau)R_{XX}(\tau)$	$S_{XX}(\omega-\omega_0)$
$X(t)\cos(\omega_0 t+\Theta)$ with independent Θ uniform on $[-\pi,+\pi]$	$\frac{1}{2}R_{XX}(\tau)\cos(\omega_0\tau)$	$\frac{1}{4}[S_{XX}(\omega+\omega_0)+S_{XX}(\omega-\omega_0)]$
$X(t)+b\ \ (E[X(t)]=0)$	$R_{XX}(\tau)+\|b\|^2$	$S_{XX}(\omega)+2\pi\|b\|^2\delta(\omega)$

In operator notation we have,

$$S_{XX}=FT\{R_{XX}\}$$

and

$$R_{XX}=IFT\{S_{XX}\},$$

where FT and IFT stand for the respective Fourier operators.

The name power spectral density (psd) will be justified later. All that we have done thus far is define it as the Fourier transform of $R_{XX}(\tau)$. We can also define the Fourier transform of the cross-correlation function $R_{XY}(\tau)$ to obtain a frequency function called the *cross-power spectral density*,

$$S_{XY}(\omega)\triangleq\int_{-\infty}^{+\infty}R_{XY}(\tau)e^{-j\omega\tau}\,d\tau. \tag{9.5-8}$$

We will see later that the psd $S_{XX}(\omega)$, is real and everywhere nonnegative and in fact, as the name implies, has the interpretation of a density function for average power versus frequency. By contrast, the cross-power spectral density has no such interpretation and is generally complex valued.

We next list some properties of the psd $S_{XX}(\omega)$:

1. $S_{XX}(\omega)$ is real valued since $R_{XX}(\tau)$ is conjugate symmetric.
2. If $X(t)$ is a real-valued WSS process, then $S_{XX}(\omega)$ is an even function since $R_{XX}(\tau)$ is real and even. Otherwise $S_{XX}(\omega)$ may not be an even function of ω.
3. $S_{XX}(\omega)\geq 0$ (to be shown in Theorem 9.5-1).

Additional properties of the psd are shown in Table 9.5-1. One could go on to expand this table, but it will suit our purposes to stop at this point. One comment is in order: We note the simplicity of these operations in the frequency domain. This suggests that for LSI systems and stationary or WSS random processes, we should solve for output correlation

functions by first transforming the input correlation function into the frequency domain, carry out the indicated operations, and then transform back to the correlation domain. This is completely analogous to the situation in deterministic linear system theory for shift-invariant systems.

Another comment would be that if the interpretation of $S_{XX}(\omega)$ as a density of average power is correct, then the constant or mean component has all its average power concentrated at $\omega = 0$ by the last entry in the table. Also by the next-to-last two entries in the table, modulation by the frequency ω_0 shifts the distribution of average power up in frequency by ω_0. Both of these results should be quite intuitive.

Example 9.5-2

(*power spectral density of white noise*) The correlation function of a white noise process $W(t)$ with parameter σ^2 is given by $R_{WW}(\tau) = \sigma^2\delta(\tau)$. Hence the power spectral density (psd), its Fourier transform, is just

$$S_{WW}(\omega) = \sigma^2, \quad -\infty < \omega < +\infty.$$

The psd is thus flat, and hence the name, *white noise*, by analogy to white light, which contains equal power at every wavelength. Just like white light, white noise is an idealization that cannot physically occur, since as we have seen earlier $R_{WW}(0) = \infty$, necessitating infinite power. Again, we note that the parameter σ^2 must be interpreted as a power density in the case of white noise.

An Interpretation of the psd

Given a WSS process $X(t)$, consider the finite support segment,

$$X_T(t) \triangleq X(t)I_{[-T,+T]}(t),$$

where $I_{[-T,+T]}$ is an indicator function equal to 1 if $-T \le t \le +T$ and equal to 0 otherwise, and $T > 0$. We can compute the Fourier transform of X_T by the integral

$$FT\{X_T(t)\} = \int_{-T}^{+T} X(t)e^{-j\omega t}\,dt.$$

The magnitude squared of this random variable is

$$|FT\{X_T(t)\}|^2 = \int_{-T}^{+T}\int_{-T}^{+T} X(t_1)X^*(t_2)e^{-j\omega(t_1-t_2)}dt_1\,dt_2.$$

Dividing by $2T$ and taking the expectation, we get

$$\frac{1}{2T}E\left[|FT\{X_T(t)\}|^2\right] = \frac{1}{2T}\int_{-T}^{+T}\int_{-T}^{+T} R_{XX}(t_1-t_2)e^{-j\omega(t_1-t_2)}\,dt_1\,dt_2, \quad (9.5\text{-}9a)$$

To evaluate the double integral on the right, introduce the new coordinate system $s = t_1 + t_2, \tau = t_1 - t_2$. The relationship between the (s,τ) and (t_1,t_2) coordinate systems

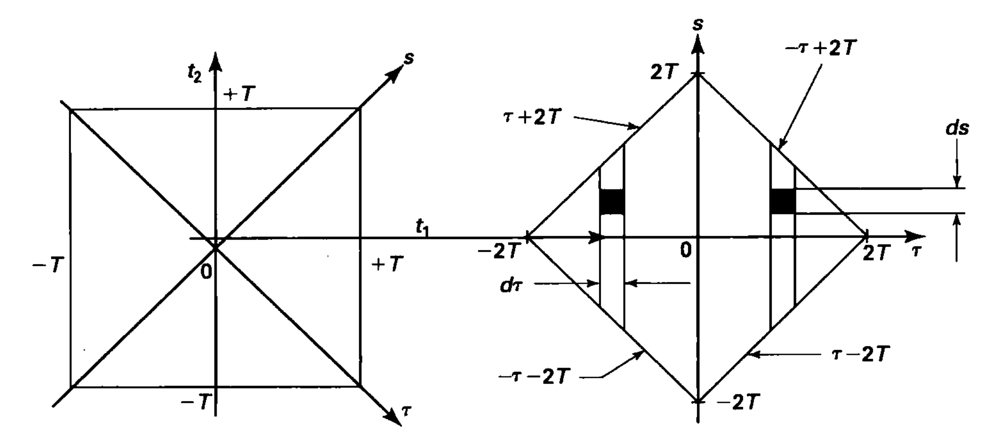


Figure 9.5-1 (a) Square region in (t_1, t_2) plane; (b) integration in diamond-shaped region created by the transformation $s = t_1 + t_2, \tau = t_1 - t_2$.

is shown in Figure 9.5-1a. The Jacobian (scale-change) of this transformation is 1/2 and the region of integration is the diamond-shaped surface $\wp$ shown in Figure 9.5-1b, which is Figure 9.5-1a rotated counterclockwise 45° and whose sides have length $T\sqrt{2}$. The double integral in Equation 9.5-9a then becomes

$$\frac{1}{4T}\iint_{\wp} R_{XX}(\tau)e^{-j\omega\tau}\,d\tau\,ds$$

$$= \frac{1}{4T}\left\{\int_{-2T}^{0} R_{XX}(\tau)e^{-j\omega\tau}\left[\int_{-(2T+\tau)}^{2T+\tau} ds\right]d\tau\right\}$$

$$+\frac{1}{4T}\left\{\int_{0}^{2T} R_{XX}(\tau)e^{-j\omega\tau}\left[\int_{-(2T-\tau)}^{2T-\tau} ds\right]d\tau\right\} = \int_{-2T}^{+2T}\left[1-\frac{|\tau|}{2T}\right]R_{XX}(\tau)e^{-j\omega\tau}d\tau.$$

In the limit as $T \to +\infty$, this integral tends to Equation 9.5-6 for an integrable R_{XX}; thus

$$S_{XX}(\omega) = \lim_{T\to\infty}\frac{1}{2T}E\left[|FT\{X_T(t)\}|^2\right] \tag{9.5-9b}$$

so that $S_{XX}(\omega)$ is real and nonnegative and is related to average power at frequency ω.

We next look at two examples of the computation of psd's corresponding to correlation functions we have seen earlier.

Example 9.5-3

Find the power spectral density for the following exponential autocorrelation function with parameter $\alpha > 0$:

$$R_{XX}(\tau) = \exp(-\alpha|\tau|), \quad -\infty < \tau < +\infty.$$

This is the autocorrelation function of the random telegraph signal (RTS) discussed in Section 9.2. Its psd is computed as

$$
\begin{aligned}
S_{XX}(\omega) &= \int_{-\infty}^{+\infty} R_{XX}(\tau)e^{-j\omega\tau}d\tau = \int_{-\infty}^{+\infty} e^{-\alpha|\tau|}e^{-j\omega\tau}\,d\tau \\
&= \int_{-\infty}^{0} e^{(\alpha-j\omega)\tau}\,d\tau + \int_{0}^{\infty} e^{-(\alpha+j\omega)\tau}\,d\tau \\
&= 2\alpha/[\alpha^2+\omega^2], \quad -\infty < \omega < +\infty.
\end{aligned}
$$

This function is plotted in Figure 9.5-2 for $\alpha = 3$. We see that the peak value is at the origin and equal to $2/\alpha$. The "bandwidth" of the process is seen to be α on a 3 dB basis (if S_{XX} is indeed a power density, to be shown). We note that while there is a cusp at the origin of the correlation function R_{XX}, there is no cusp in its spectral density S_{XX}. In fact S_{XX} is continuous and differentiable everywhere. (It is true that S_{XX} will always be continuous if R_{XX} is absolutely integrable.)

Figure 9.5-2 was created using MATLAB with the short m-file:

```
clear alpha=3;
b = [1.0 0.0 alpha^2];
w = linspace(-10,+10);
den = polyval(b,w);
num = 2*alpha;
S = num./den;
plot (w,S)
```

We note that the psd decays rather slowly, and thus the RTS process requires a significant amount of bandwidth. The reason the tails of the psd are so long is due to the jumps in the RTS sample functions.

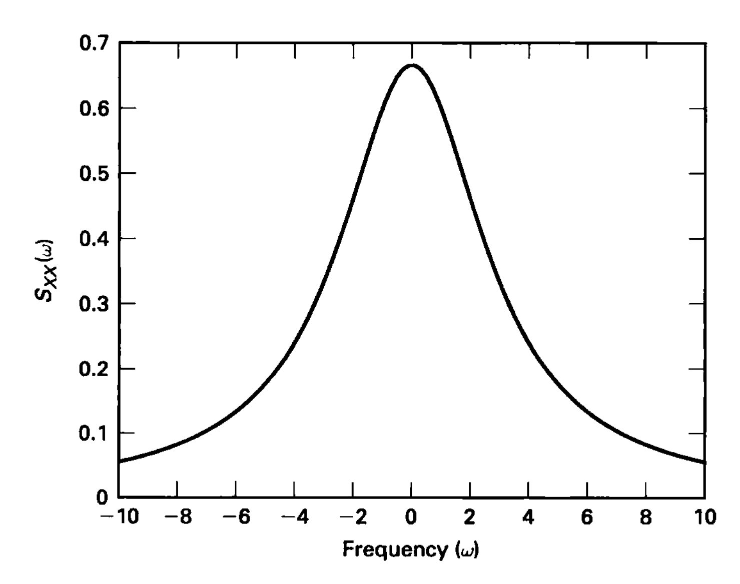


Figure 9.5-2 Plot of psd for exponential autocorrelation function.

Example 9.5-4

(*psd of triangular autocorrelation*) Consider an autocorrelation function that is triangular in shape such that the correlation goes to zero at shift $T > 0$,

$$R_{XX}(\tau) = \max\left[1 - \frac{|\tau|}{T}, 0\right].$$

One way this could arise is the asynchronous binary signaling (ABS) process introduced in Section 9.2. This function is plotted as Figure 9.5-3. If we realize that this triangle can be written as the convolution of two rectangular pulses, each of width T and height $1/\sqrt{T}$, then we can use the convolution theorem of the Fourier transform [9-3,-4] to see that the psd of the triangular correlation function is just the square of the Fourier transform of the rectangular pulse, that is, the *sinc* function. The transform of the rectangular pulse is

$$\sqrt{T}\frac{\sin(\omega T/2)}{(\omega T/2)},$$

and the power spectral density S_{XX} of the triangular correlation function is thus

$$S_{XX}(\omega) = T\left(\frac{\sin(\omega T/2)}{\omega T/2}\right)^2. \tag{9.5-10}$$

As a check we note that $S_{XX}(0)$ is just the area under the correlation function, that in the triangular case is easily seen to be T. Thus checking,

$$S_{XX}(0) = \int_{-\infty}^{+\infty} R_{XX}(\tau)\, d\tau = 2 \cdot \frac{1}{2} \cdot 1 \cdot T.$$

Another way the triangular correlation function can arise is the running integral average operating on white noise. Consider

$$X(t) \triangleq \frac{1}{\sqrt{T}} \int_{t-T}^{t} W(\tau)\, d\tau,$$

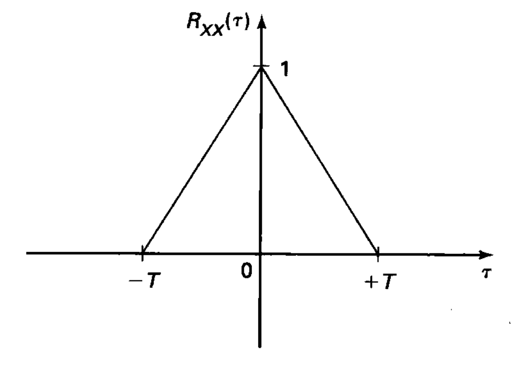


Figure 9.5-3 A triangular autocorrelation function.

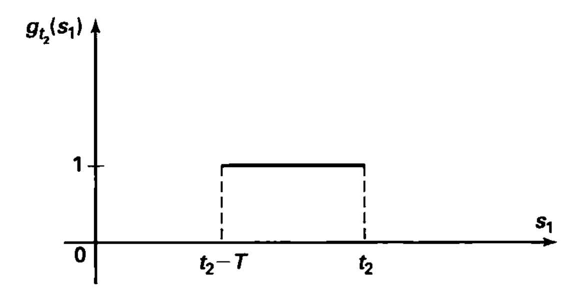


Figure 9.5-4 Plot of equation versus s_1 for $t_2 > T$.

with $W(t)$ a white noise with zero mean and correlation function $R_{WW}(\tau) = \delta(\tau)$. Then $\mu_X(t) = 0$ and $E[X(t_1)X(t_2)]$ can be computed as

$$
\begin{aligned}
R_{XX}(t_1, t_2) &= \frac{1}{T} \int_{t_1-T}^{t_1} \int_{t_2-T}^{t_2} R_{WW}(s_1 - s_2) ds_1 \, ds_2 \\
&= \frac{1}{T} \int_{t_1-T}^{t_1} \left[\int_{t_2-T}^{t_2} \delta(s_2 - s_1) \, ds_2 \right] ds_1.
\end{aligned}
$$

Now defining the inner integral as

$$
g_{t_2}(s_1) \triangleq \int_{t_2-T}^{t_2} \delta(s_2 - s_1) ds_2 = \begin{cases} 1, & t_2 - T \leq s_1 \leq t_2, \\ 0, & \text{else,} \end{cases}
$$

which as a function of s_1 looks as shown in Figure 9.5-4, so

$$
\begin{aligned}
R_{XX}(t_1, t_2) &= \frac{1}{T} \int_{t_1-T}^{t_1} g_{t_2}(s_1) ds_1 \\
&= \max \left[1 - \frac{|t_1 - t_2|}{T}, 0 \right].
\end{aligned}
$$

More on White Noise

The correlation function of white noise is an impulse (Equation 9.3-6), so its psd is a constant

$$
S_{WW}(\omega) = \sigma^2, \qquad -\infty < \omega < +\infty.
$$

The name white noise thus arises out of the fact that the power spectral density is constant at all frequencies just as in white light, which contains all wavelengths in equal amounts.† Here we look at the white noise process as a limit approached by a sequence of second-order

†A mathematical idealization! Physics tells us that, for realistic models, the power density must tend toward zero as $\omega \to \infty$.

processes. To this end consider an independent increment process (cf. Definition 9.2-1) with zero mean such as the Wiener process ($R_{XX}(t_1, t_2) = \sigma^2 \min(t_1, t_2)$) or a centered Poisson process, that is, $N_c(t) = N(t) - \lambda t$, with correlation $R_{N_c N_c}(t_1, t_2) = \lambda \min(t_1, t_2)$. Actually we need only uncorrelated increments here; thus we require $X(t)$ only to have uncorrelated increments. For such processes we have by Equation 9.2-17,

$$E\left[(X(t+\Delta) - X(t))^2\right] = \alpha\Delta,$$

where α is the variance parameter.

Thus upon letting $X_\Delta(t)$ denote the first-order difference divided by Δ,

$$X_\Delta(t) \triangleq [X(t+\Delta) - X(t)]/\Delta,$$

we have

$$E[X_\Delta^2(t)] = \alpha/\Delta$$

and

$$E[X_\Delta(t_1)X_\Delta(t_2)] = 0 \quad \text{for} \quad |t_2 - t_1| > \Delta.$$

If we consider $|t_2 - t_1| < \Delta$, we can do the following calculation, which shows that the resulting correlation function is triangular, just as in Example 9.5-4. Since $X(t_1+\Delta) - X(t_1)$ is distributed as $N(0, \Delta)$, taking $t_1 < t_2$ and shifting t_1 to 0, and t_2 to $t_2 - t_1$, the expectation becomes

$$\begin{aligned}
&\frac{1}{\Delta^2} E\left[X(\Delta)\left(X(t_2 - t_1 + \Delta) - X(t_2 - t_1)\right)\right] \\
&\quad = \frac{1}{\Delta^2} E\left[X(\Delta)\left(X(\Delta) - X(t_2 - t_1)\right)\right] \quad \text{since } (\Delta, t_2 - t_1 + \Delta] \cap (0, \Delta] = \phi, \\
&\quad = \frac{1}{\Delta^2}[\alpha\Delta - \alpha(t_2 - t_1)] = \frac{\alpha}{\Delta}[1 - (t_2 - t_1)/\Delta].
\end{aligned}$$

Thus, the process generated by the first-order difference is WSS (the mean is zero) and has correlation function $R_{\Delta\Delta}(\tau)$ given as

$$R_{\Delta\Delta}(\tau) = \frac{\alpha}{\Delta} \max\left[1 - \frac{|\tau|}{\Delta}, 0\right].$$

We note from Figure 9.5-5 that as Δ goes to zero this correlation function tends to a delta function.

Since we just computed the Fourier transform of a triangular function in Example 9.5-4, we can write the psd by inspection as

$$S_{\Delta\Delta}(\omega) = \alpha \left(\frac{\sin(\omega\Delta/2)}{\omega\Delta/2}\right)^2.$$

This psd is approximately flat out to $|\omega| = \pi/(3\Delta)$. As $\Delta \to 0$, $S_{\Delta\Delta}(\omega)$ approaches the constant α everywhere. Thus as $\Delta \to 0$, $X_\Delta(t)$ "converges" to white noise, the *formal* derivative of an uncorrelated increments process,

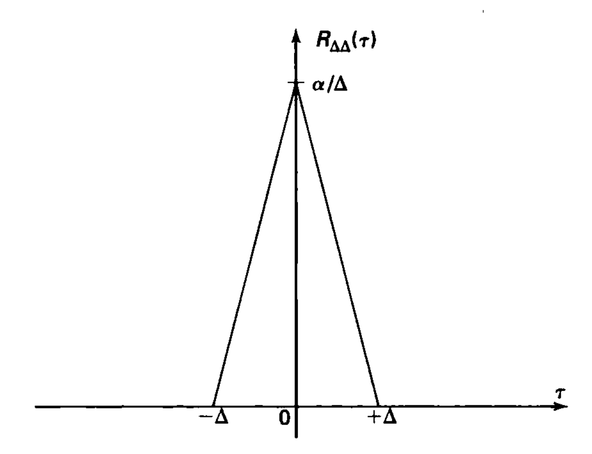


Figure 9.5-5 Correlation function of $X_\Delta(t)$.

$$\begin{aligned} R_{\dot{X}\dot{X}}(t_1, t_2) &= \frac{\partial^2}{\partial t_1 \partial t_2}[\sigma^2 \min(t_1, t_2)] \\ &= \frac{\partial}{\partial t_1}[\sigma^2 u(t_1 - t_2)] \\ &= \sigma^2 \delta(t_1 - t_2). \end{aligned}$$

If one has a system that is continuous in its response to stimuli, then we say that the *system is continuous*; that is, the system operator is a continuous operator. This would mean, for example, that the output would change only slightly if the input changed slightly. A stable differential or difference equation is an example of such a continuous operator. We will see that for linear shift-invariant systems that are described by system functions, the response to the random process $X_\Delta(t)$ will change only slightly when Δ changes, if Δ is small and if the systems are lowpass in the sense that the system function tends to zero as $|\omega| \to +\infty$. Thus the white noise can be seen as a convenient artifice for more easily constructing this limiting output. (See Problem 9.36.)

If we take Fourier transforms of both sides of Equation 9.5-3 we obtain the *cross-power spectral density*,

$$S_{YX}(\omega) = H(\omega) S_{XX}(\omega). \tag{9.5-11}$$

Since S_{YX} is a frequency-domain representation of the cross-correlation function R_{YX}, Equation 9.5-11 tells us that $Y(t)$ and $X(t)$ will have high cross correlation at those frequencies ω where the product of $H(\omega)$ and $S_{XX}(\omega)$ is large. Similarly, from Equation 9.5-5, we can obtain

$$S_{XY}(\omega) = H^*(\omega) S_{XX}(\omega). \tag{9.5-12}$$

From the fundamental Equation 9.5-4, repeated here for convenience,

$$R_{YY}(\tau) = h(\tau) * R_{XX}(\tau) * h^*(-\tau), \tag{9.5-13}$$

we get, upon Fourier transformation, in the spectral density domain,

$$S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega) = G(\omega) S_{XX}(\omega). \tag{9.5-14}$$

These two equations are among the most important in the theory of stationary random processes. In particular, Equation 9.5-14 shows how the average power in the output process is composed solely as the average input power at that frequency multiplied by $|H(\omega)|^2$, the power gain of the LSI system. We can call $G(\omega) = |H(\omega)|^2$ the *psd transfer function.*

Example 9.5-5

(*average power*) The transfer function of an LSI system is given by

$$H(\omega) = \text{sgn}(\omega)\left(\frac{\omega}{2\pi}\right)^2 \exp\left[-j\left(\omega \cdot \frac{8}{\pi}\right)\right] W(\omega),$$

where sgn$(\cdot)$ is the algebraic sign function, and where the frequency window function

$$W(\omega) \triangleq \begin{cases} 1, & \text{for } |\omega| \leq 40\pi \\ 0, & \text{else.} \end{cases}$$

Let the WSS input random process have autocorrelation function,

$$R_{XX}(\tau) = \frac{5}{2}\delta(\tau) + 2.$$

Compute the average measurable power in the band 0.0 to 1.0 Hertz (single-sided). In radians, this is the double-sided range -2π to 2π. First we Fourier transform $R_{XX}(\tau)$ to obtain $S_{XX}(\omega) = \frac{5}{2} + 4\pi\delta(\omega)$. Next we compute the psd transfer function $G(\omega) = |H(\omega)|^2 = \left(\frac{\omega}{2\pi}\right)^4 W(\omega)$. The output psd then is

$$S_{YY}(\omega) = \frac{5}{2}\left(\frac{\omega}{2\pi}\right)^4 W(\omega),$$

and the total average output power would be calculated as

$$R_{YY}(0) = \frac{1}{2\pi}\int_{-40\pi}^{+40\pi} \frac{5}{2}\left(\frac{\omega}{2\pi}\right)^4 d\omega,$$

while the power in the band $[-2\pi, +2\pi]$ is

$$P = \frac{1}{2\pi}\int_{-2\pi}^{+2\pi} \frac{5}{2}\left(\frac{\omega}{2\pi}\right)^4 d\omega$$

$$= 1 \text{ watt}.$$

The following comment on Equations 9.5-3 through 9.5-14 may help you keep track of the conjugates and minus signs. Notice that the conjugate and negative argument on

the impulse response, which becomes simply a conjugate in the frequency domain, arises in connection with the second factor in the correlation. The $h(\tau)$ without the conjugate or negative time argument comes from the linear operation implied by the first subscript, that is, the first factor in the correlation.

With reference to Equation 9.5-11 we see that the cross-spectral density function can be complex and hence has no positivity or conjugate symmetry properties, since those that S_{XX} has will be lost upon multiplication with an arbitrary, generally complex H. On the other hand, as shown in Equation 9.5-14, the psd of the output will share the real and nonnegative aspects of the psd of the input, since multiplication with $|H|^2$ will not change these properties. Table 9.5-2 sets forth all the above relations for easy reference.

We are now in a position to show that the psd $S(\omega)$ has a precise interpretation as a density for average power versus frequency. We will show directly that $S(\omega) \geq 0$ for all ω and that the average power in the frequency band (ω_1, ω_2) is given by the integral of $S(\omega)$ over that frequency band.

Theorem 9.5-1 Let $X(t)$ be a stationary, second-order random process with correlation function $R_{XX}(\tau)$ and power spectral density $S_{XX}(\omega)$. Then $S_{XX}(\omega) \geq 0$ and for all $\omega_2 \geq \omega_1$,

$$\frac{1}{2\pi}\int_{\omega_1}^{\omega_2} S_{XX}(\omega)d\omega$$

is the average power in the frequency band (ω_1, ω_2).

Proof Let $\omega_2 > \omega_1$ both be real numbers. Define a filter transfer function as follows:

$$H(\omega) \triangleq \begin{cases} 1, & \omega \in (\omega_1, \omega_2) \\ 0, & \text{else}, \end{cases}$$

Table 9.5-2 Input/Output Relations for Linear Systems with WSS Inputs

WSS Random Process:	Output Mean:
$Y(t) = h(t) * X(t)$	$\mu_Y = H(0)\mu_X$
Crosscorrelations:	Cross–Power Spectral Densities:
$R_{XY}(\tau) = R_{XX}(\tau) * h^*(-\tau)$	$S_{XY}(\omega) = S_{XX}(\omega)H^*(\omega)$
$R_{YX}(\tau) = h(\tau) * R_{XX}(\tau)$	$S_{YX}(\omega) = H(\omega)S_{XX}(\omega)$
$R_{YY}(\tau) = R_{YX}(\tau) * h^*(-\tau)$	$S_{YY}(\omega) = S_{YX}(\omega)H^*(\omega)$
Autocorrelation:	Power Spectral Density:
$R_{YY}(\tau) = h(\tau) * R_{XX}(\tau) * h^*(-\tau)$	$S_{YY}(\omega) = \|H(\omega)\|^2 S_{XX}(\omega)$
$= g(\tau) * R_{XX}(\tau)$	$= G(\omega)S_{XX}(\omega)$

Output Power and Variance:

$$E\{|Y(t)|^2\} = R_{YY}(0) = \tfrac{1}{2\pi}\int_{-\infty}^{+\infty} |H(\omega)|^2 S_{XX}(\omega)d\omega$$

$$\sigma_Y^2 = R_{YY}(0) - |\mu_Y|^2$$

and note that it passes signals only in the band (ω_1, ω_2). If $X(t)$ is input to this filter, the psd of the output $Y(t)$ is (by Equation 9.5-14)

$$S_{YY}(\omega) = \begin{cases} S_{XX}(\omega), & \omega \in (\omega_1, \omega_2) \\ 0, & \text{else.} \end{cases}$$

Now the output power in $Y(t)$ has average value $E[|Y(t)|^2] = R_{YY}(0)$,

$$R_{YY}(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S_{YY}(\omega)d\omega = \frac{1}{2\pi}\int_{\omega_1}^{\omega_2} S_{XX}(\omega)d\omega \geq 0,$$

and this holds for all $\omega_2 > \omega_1$. So by choosing $\omega_2 \simeq \omega_1$ we can conclude that $S_{XX}(\omega) \geq 0$ for all ω and that the function S_{XX} thus has the interpretation of a power density in the sense that if we integrate this function across a frequency band, we get the average power in that band. ■

We saw earlier that the conditions that a function must meet to be a valid correlation or covariance function are rather strong. In fact, we have seen that the function must be positive semidefinite, although we have not in fact shown that this condition is sufficient. It turns out that one more advantage of working in the frequency domain is the ease with which we can specify when a given frequency function qualifies as a power spectral density. The function simply must be real and nonnegative, that is, $S(\omega) \geq 0$. We can see this for a given function $F(\omega) \geq 0$ by taking a filter with transfer function $H(\omega) = \sqrt{F(\omega)}$ and letting the input be white noise with $S_{WW} = 1$. Then by Equation 9.5-14 the output psd is $S_{XX}(\omega) = F(\omega)$, thus showing that F is a valid psd. If the random process is real valued, as it most often is, then we also need $F(\omega)$ to be an even function to satisfy psd property (2) listed just after Definition 9.5-1. All this can be formalized as follows.

Theorem 9.5-2 Let $F(\omega)$ be an integrable function that is real and nonnegative; that is, $F(\omega) \geq 0$ for all ω. Then there exists a stationary random process with power spectral density $S(\omega) = F(\omega)$. If the random process is to be real valued, then $F(\omega)$ must be an even function of ω. ■

We now see that the test for a valid spectral density function is much easier than the condition of positive semidefiniteness for the correlation function. In fact, it is relatively easy to show that the positive semidefinite condition on a function is equivalent to the nonnegativity of its Fourier transform, and hence that positive semidefiniteness is the sufficient condition for a function to be a valid correlation or covariance function. First, by Theorem 9.5-2 we know that the positive semidefinite condition is implied by the nonnegativity of $S(\omega)$. To show equivalence, it remains to show that the positive semidefinite condition on a function $f(\tau)$ implies that its Fourier transform $F(\omega)$ is nonnegative. We proceed as follows: Since $f(\tau)$ is positive semidefinite we have,

$$\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m^* f(\tau_n - \tau_m) \geq 0.$$

Also since

$$f(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) e^{+j\omega\tau} d\omega,$$

we have

$$\frac{1}{2\pi} \sum_n \sum_m \left(a_n a_m^* \int_{-\infty}^{+\infty} F(\omega) e^{+j\omega(\tau_n - \tau_m)} \right) d\omega \geq 0,$$

which can be rewritten as

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) \left[\sum_n \sum_m a_n a_m^* e^{+j\omega(\tau_n - \tau_m)} \right] d\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega) \left| \sum_{n=1}^{N} a_n e^{+j\omega\tau_n} \right|^2 d\omega \geq 0,$$

where we recognize the term inside the magnitude square sign as a so-called transversal or tapped delay-line filter. Thus by choosing N large enough, with the τ_n equally spaced, we can select the a_n's to arbitrarily approximate any ideal filter transfer function $H(\omega)$. Then by choosing H to be very narrow bandpass filters centered at each value of ω, we can eventually conclude that $F(\omega) \geq 0$ for all ω, $-\infty < \omega < +\infty$. We have thereby established the following theorem.

Theorem 9.5-3 A necessary and sufficient condition for $f(\tau)$ to be a correlation function is that it be positive semidefinite. ■

Incidentally, there is an analogy here for probability density functions, which can be regarded as the Fourier transforms of their CFs. As we know, nonnegativity is the sufficient condition for a function to be a valid pdf (assuming that it is normalized to integrate to one); thus the probability density is analogous to the power spectral density; and in fact one can define a spectral distribution function [9–7] analogous to the cumulative distribution function. Thus the CF and the correlation function are also analogous and so both must be positive semidefinite to be valid for their respective roles. Also for the CF the normalization of the probability density to integrate to one imposes the condition $\Phi(0) = 1$, which is easily met by scaling an arbitrary positive semidefinite function that is not identically zero.

Stationary Processes and Differential Equations

We shall now examine stochastic differential equations with a stationary or at least WSS input, and also with the linear constant-coefficient differential equation (LCCDE) valid for all time. We assume that the equation is stable in the bounded-input, bounded-output (BIBO) sense, so that the resulting output process is also stationary (or WSS if that is the condition on the input process).

Thus consider the following general LCCDE:

$$\begin{aligned} &a_N Y^{(N)}(t) + a_{N-1} Y^{(N-1)}(t) + \ldots + a_0 Y(t) \\ &\quad = b_M X^{(M)}(t) + b_{M-1} X^{(M-1)}(t) + \ldots + b_0 X(t), \qquad -\infty < t < +\infty. \end{aligned}$$

This represents the relationship between output $Y(t)$ and input $X(t)$ in a linear system with frequency response

$$H(\omega) = B(\omega)/A(\omega), \quad \text{with } a_0 \neq 0,$$

where

$$B(\omega) \triangleq \sum_{m=0}^{M} b_m (j\omega)^m$$

and

$$A(\omega) \triangleq \sum_{n=0}^{N} a_n (j\omega)^n,$$

which is a *rational function* with *numerator polynomial* $B(\omega)$ and *denominator polynomial* $A(\omega)$. Because the system is stable, we can apply the results of the previous section to obtain

$$\mu_Y = \mu_X H(0)$$
$$S_{YX}(\omega) = H(\omega)\, S_{XX}(\omega),$$

and

$$S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega),$$

where

$$H(0) = b_0/a_0 \quad \text{and} \quad |H(\omega)|^2 = |B(\omega)|^2/|A(\omega)|^2.$$

So

$$\mu_Y = (b_0/a_0)\, \mu_X \quad \text{and} \quad S_{YY}(\omega) = \left(|B(\omega)|^2/|A(\omega)|^2\right) S_{XX}(\omega).$$

This frequency-domain analysis method is generally preferable to the time-domain approach but is restricted to the case where both the input and output processes are at least WSS. After we obtain the various spectral densities, then we can use the *IFT* to obtain the correlation and covariance functions if they are desired. The calculation of the required *IFTs* is often easier if viewed as an inverse two-sided *Laplace transform.* The Laplace transform of Equation 9.5-3 is

$$\mathsf{S}_{YX}(s) = \mathsf{H}(s)\mathsf{S}_{XX}(s) \tag{9.5-15}$$

while the Laplace transform of Equation 9.5-13 is written

$$\mathsf{S}_{YY}(s) = \mathsf{H}(s)\mathsf{H}(-s)\mathsf{S}_{XX}(s) \tag{9.5-16}$$

in light of $h^*(-\tau) \leftrightarrow \mathsf{H}(-s)$. Recalling the definition of the two-sided Laplace transform [9–3], for any $f(\tau)$

$$\mathsf{F}(s) \triangleq \int_{-\infty}^{+\infty} f(\tau) e^{-s\tau} d\tau,$$

we note that such a function of the complex variable s may be obtained from the Fourier transform $F(\omega)$, a function of the real variable ω, by a two-step procedure. First set

$$\mathsf{F}(s)_{s=j\omega} \triangleq F(\omega)$$

and then replace $j\omega$ by s. An analogous extension method was used earlier for the discrete-time case in Chapter 8 where the Fourier transform was extended to the entire complex plane by the Z-transform.

Example 9.5-6

(*output correlation—first-order system*) Consider the first-order differential equation

$$Y'(t) + \alpha Y(t) = X(t), \qquad \alpha > 0,$$

with stationary input $X(t)$ with mean $\mu_X = 0$ and impulse covariance function $K_{XX}(\tau) = \delta(\tau)$. The system function is easily seen to be

$$H(\omega) = \frac{1}{\alpha + j\omega},$$

and the psd of the input process is

$$S_{XX}(\omega) = 1,$$

so we have the following cross- and output-power spectral densities:

$$S_{YX}(\omega) = H(\omega)S_{XX}(\omega) = \frac{1}{\alpha + j\omega},$$

$$S_{YY}(\omega) = |H(\omega)|^2 S_{XX}(\omega) = \frac{1}{|\alpha + j\omega|^2} = \frac{1}{\alpha^2 + \omega^2}.$$

We now convert to Laplace transforms, with $s = j\omega$,

$$\mathsf{S}_{YY}(j\omega) = \frac{1}{\left(\alpha^2 - (j\omega)^2\right)}$$

$$= \frac{1}{(\alpha + j\omega)(\alpha - j\omega)}$$

so that

$$\mathsf{S}_{YY}(s) = \frac{1}{(s + \alpha)(-s + \alpha)}.$$

Using the residue method (cf. Appendix A) or partial fraction expansion, one can then directly obtain the following output correlation function by inverse Laplace transform:

$$R_{YY}(\tau) = \frac{1}{2\alpha}\exp(-\alpha|\tau|), \qquad -\infty < \tau < +\infty,$$

which is also the output covariance function since $\mu_Y = 0$. By the above equation for $S_{YX}(\omega)$ we also obtain the cross-correlation function $R_{YX}(\tau) = \exp(-\alpha\tau)u(\tau)$.

In Example 9.5-6 it is interesting that $R_{YX}(\tau)$ is 0 for $\tau < 0$. This means that the output Y is orthogonal to all future values of the input X, a white noise in this case. This occurs because of two reasons: The system is causal and the input is a white noise process. The system causality requires that the output not depend *directly* on (i.e., not be a function of) future inputs but only depend directly on present and past inputs. The whiteness of the input X guarantees that the past and present inputs will be uncorrelated with future inputs. Combining both conditions we see that there will be no cross-correlation between the present output and the future inputs. If we assume additionally that the input is Gaussian, then the input process is an independent process and the output becomes independent of all future inputs. Then we can say that the causality of the system prevents the *direct dependence* of the present output on future inputs, and the independent process input prevents any *indirect dependence.* These concepts are important to the theory of Markov processes as used in estimation theory (cf. Chapter 11).

Example 9.5-7

(*output correlation function—second-order system*) Consider the following second-order LCCDE:

$$\frac{d^2Y(t)}{dt^2} + 3\frac{dY(t)}{dt} + 2Y(t) = 5X(t),$$

again with white noise input as in the previous example. Here the system function is

$$H(\omega) = \frac{5}{(j\omega)^2 + 3j\omega + 2} = \frac{5}{(2-\omega^2) + j3\omega}.$$

Thus analogously to Example 9.5-6 the output psd becomes

$$S_{YY}(\omega) = \frac{25}{(2-\omega^2)^2 + (3\omega)^2} = \frac{25}{\omega^4 + 5\omega^2 + 4}.$$

Applying the residue method to evaluate the IFT, we define the function of a complex variable $\mathsf{S}_{YY}(s)|_{s=j\omega} \triangleq S_{YY}(\omega)$ and rewrite the right-hand side in terms of the complex variable $j\omega$ to obtain

$$\mathsf{S}_{YY}(j\omega) = \frac{25}{(j\omega)^4 - 5(j\omega)^2 + 4}.$$

Substituting $s = j\omega$, we get

$$\mathsf{S}_{YY}(s) = \frac{25}{s^4 - 5s^2 + 4},$$

which factors as

$$\frac{5}{(s+2)(s+1)} \cdot \frac{5}{(-s+2)(-s+1)} = \mathsf{H}(s)\mathsf{H}(-s),$$

where $\mathsf{H}(s)$ is the Laplace transform system function. Then the inverse Laplace transform yields the output correlation function

$$R_{YY}(\tau) = 25\left[\frac{1}{6}\exp(-|\tau|) - \frac{1}{12}\exp(-2|\tau|)\right], \qquad -\infty < \tau < +\infty.$$

We leave the details of the calculation to the interested reader.

9.6 PERIODIC AND CYCLOSTATIONARY PROCESSES

Besides stationarity and its wide-sense version, two other classes of random processes are often encountered. They are periodic and cyclostationary processes and are here defined.

Definition 9.6-1 A random process $X(t)$ is *wide-sense periodic* if there is a $T > 0$ such that

$$\mu_X(t) = \mu_X(t+T) \qquad \text{for all } t$$

and

$$K_{XX}(t_1,t_2) = K_{XX}(t_1+T,t_2) = K_{XX}(t_1,t_2+T) \qquad \text{for all } t_1, t_2.$$

The smallest such T is called the *period.* Note that $K_{XX}(t_1,t_2)$ is then periodic with period T along both axes. ■

An example of a wide-sense periodic random process is the random complex exponential of Example 9.4-1. In fact, the random Fourier series representation of the process

$$X(t) = \sum_{k=1}^{\infty} A_k \exp\left(j\frac{2\pi kt}{T}\right) \tag{9.6-1}$$

with random variable coefficients A_k would also be wide-sense periodic. A wide-sense periodic process can also be WSS, in which case we call it *wide-sense periodic stationary.* We will consider these processes further in Chapter 10, where we also refer to them as *mean-square periodic.* The covariance function of a wide-sense periodic process is generically sketched in Figure 9.6-1. We see that $K_{XX}(t_1,t_2)$ is doubly periodic with a two-dimensional period of (T,T). In Chapter 10 we will see that the sample functions of a wide-sense periodic random process are periodic with probability 1, that is,

$$X(t) = X(t+T) \quad \text{for all } t,$$

except for a set of outcomes, i.e. an event, of probability zero.

Another important classification is *cyclostationarity.* It is only partially related to periodicity and is often confused with it. The reader should carefully note the difference in the following definition. Roughly speaking, cyclostationary processes have *statistics* that are periodic, while periodic processes have *sample functions* that are periodic.

Definition 9.6-2 A random process $X(t)$ is wide-sense cyclostationary if there exists a positive value T such that

$$\mu_X(t) = \mu_X(t+T) \quad \text{for all } t$$

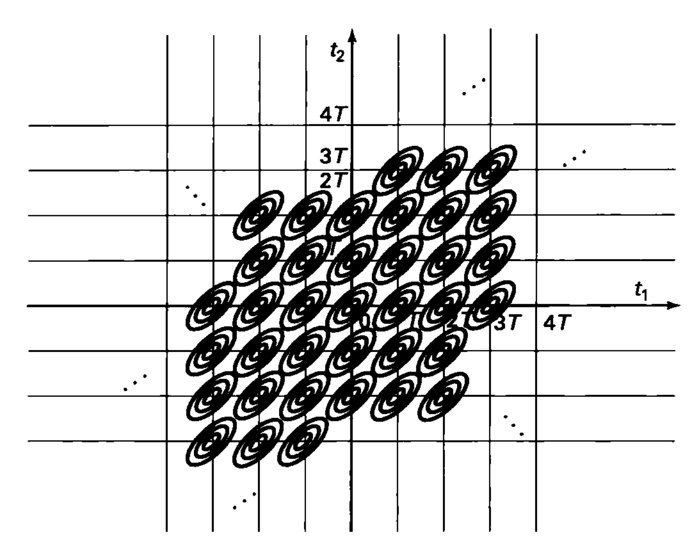


Figure 9.6-1 Possible contours of the covariance function of a wide-sense (WS) periodic random process.

and

$$K_{XX}(t_1, t_2) = K_{XX}(t_1 + T, t_2 + T) \quad \text{for all } t_1 \text{ and } t_2. \quad \blacksquare$$

An example of cyclostationarity is the random PSK process of Equation 9.2-11. Its mean function is zero and hence trivially periodic. Its covariance function (Equation 9.2-13) is invariant to a shift by T in *both* its arguments. Note that Equation 9.2-13 is *not* doubly–periodic since $R_{XX}(0, T) = 0 \neq R_{XX}(0, 0)$. Also note that the sample functions of $X(t)$ are *not periodic* in any sense.

The constant-value contours of the covariance function of a typical cyclostationary random process are shown in Figure 9.6-2. Note the difference between this configuration and that of a periodic random process, as shown in Figure 9.6-1. Effectively, cyclostationarity means that the statistics are periodic, but the process itself is not periodic.

By averaging along 45° lines (i.e., $t_1 = t_2$), we can get the WSS versions of both types of processes. The contours of constant density of the periodic process then become the straight lines of the WSS periodic process shown in Figure 9.6-3. The WSS version of a cyclostationary process just becomes an ordinary WSS process, because of the lack of any periodic structure along 135° (anti-diagonal) lines (i.e., $t_1 = -t_2$).

In addition to modulators, scanning sensors tend to produce cyclostationary processes. For example, the line-by-line scanning in television transforms the random image field into a one-dimensional random process that has been modeled as cyclostationary. In communications, cyclostationarity often arises due to waveform repetition at the baud or symbol rate.

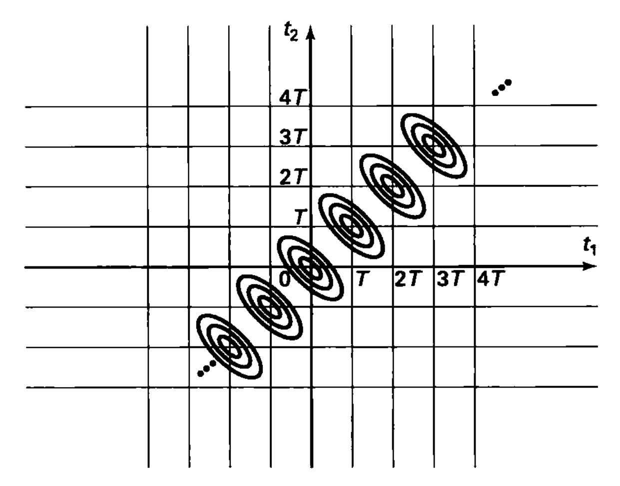


Figure 9.6-2 Possible contour plot of covariance function of WS cyclostationary random process.

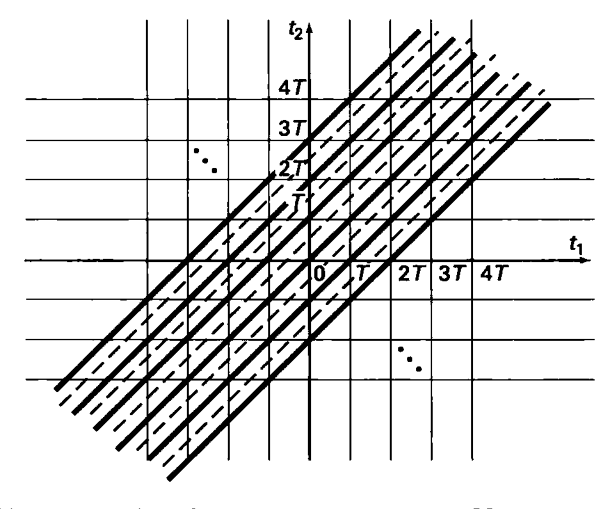


Figure 9.6-3 Possible contour plot of covariance function of WSS periodic random process. (Solid lines are maxima; dashed lines are minima.)

A place where cyclostationarity arises in signal processing is when a stationary random sequence is analyzed by a filter bank and subsampled. The subsequent filter bank synthesis involves upsampling and reconstruction filters. If the *subsampling* period is N, then the resulting synthesized random sequence will be cyclostationary with period N. When perfect reconstruction filters are used, then true stationarity will be achieved for the synthesized output.

While cyclostationary processes are not stationary or WSS except in trivial cases, it is sometimes appropriate to convert a cyclostationary process into a stationary process as in the following example.

Example 9.6-1

(*WSS PSK*) We have seen that the PSK process of Section 9.2 is cyclostationary and hence not WSS. This is easily seen with reference to Equation 9.2-13. This cyclostationarity arises from the fact that the analog angle process $\Theta_a(t)$ is stepwise constant and changes only at $t = nT$ for integer n. In many real situations the modulation process starts at an arbitrary time t, which in fact can be modeled as random from the viewpoint of the system designer. Thus in this practical case, the modulated signal process (Equation 9.2-11) is converted to

$$\widetilde{X}(t) = \cos\left(2\pi f_c t + \Theta_a(t) + 2\pi f_c T_0\right), \tag{9.6-2}$$

by the addition of a random variable T_0, which is uniformly distributed on $[0, T]$ and independent of the angle process $\Theta_a(t)$. It is then easy to see that the mean and covariance functions need only to be modified by an ensemble average over T_0, which by the uniformity of T_0 is just an integral over $[0, T]$. We thus obtain

$$\begin{aligned} R_{\widetilde{X}\widetilde{X}}(t_1+\tau, t_1) &= \frac{1}{T}\int_0^T R_{XX}(t_1+\tau+t, t_1+t)dt \\ &= \frac{1}{T}\int_{-\infty}^{+\infty} s_Q(t_1+t+\tau)s_Q(t_1+t)dt \\ &= \frac{1}{T}s_Q(\tau) * s_Q(-\tau), \end{aligned} \tag{9.6-3}$$

which is just a function of the shift τ. Thus $\widetilde{X}(t)$ is a WSS random process.

Example 9.6-2

(*power spectral density of PSK*) A WSS version of the random PSK signal was defined in Example 9.6-1 through an averaging process, where the average was taken over the message time or baud interval T. The resulting WSS random process $\widetilde{X}(t)$ had correlation function (Equation 9.6-3) given as

$$R_{\widetilde{X}\widetilde{X}}(\tau) = \frac{1}{T}s_Q(\tau) * s_Q(-\tau),$$

where $s_Q(\tau)$ was given as

$$s_Q(\tau) = \begin{cases} \sin(2\pi f_c \tau), & 0 \le \tau \le T, \\ 0, & \text{else}. \end{cases}$$

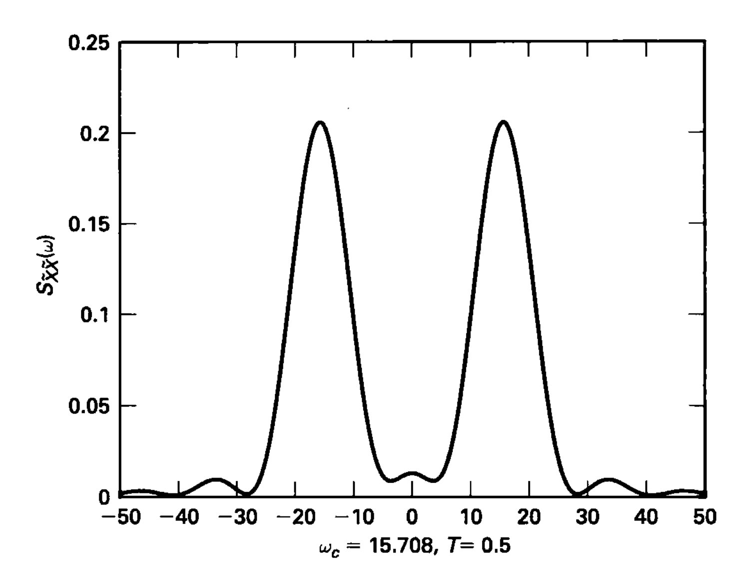


Figure 9.6-4 Power spectral density of PSK plotted for $f_c = 2.5$ and $T = 0.5$.

Then the psd of this WSS version of PSK can be calculated as

$$\begin{aligned} S_{\tilde{X}\tilde{X}}(\omega) &= FT\{R_{\tilde{X}\tilde{X}}(\tau)\} \\ &= \frac{1}{T}|FT\{s_Q(\tau)\}|^2 \\ &\approx (T/4)\left\{\left(\frac{\sin(\omega - 2\pi f_c)\frac{T}{2}}{(\omega - 2\pi f_c)\frac{T}{2}}\right)^2 + \left(\frac{\sin(\omega + 2\pi f_c)\frac{T}{2}}{(\omega + 2\pi f_c)\frac{T}{2}}\right)^2\right\}, \\ &\text{for } f_c T >> 1, \end{aligned} \tag{9.6-4}$$

which can be plotted† using MATLAB. The file **psd_PSK.m** included on this book's Web site.

Some plots were made using **psd_PSK.m**, for two different sets of values for f_c and T. First we look at the psd plot in Figure 9.6-4 for $f_c = 2.5$ and $T = 0.5$, which gives considerable overlap of the positive and negative frequency lobes of $S_{\tilde{X}\tilde{X}}(\omega)$. The lack of power concentration at the carrier frequency f_c is not surprising, since there is only a little over one period of $s_Q(t)$ in the baud interval T. The next pair of plots show a quite different case with power strongly concentrated at ω_c. This plot was computed with the values $f_c = 3.0$ and $T = 5.0$, thus giving 15 periods of the sine wave in the baud interval T. Figure 9.6-5 is a linear plot, while Figure 9.6-6 shows $S_{\tilde{X}\tilde{X}}(\omega)$ on a logarithmic scale.

---

†The reason for the approximate equals sign is that we have neglected the cross-term in Equation 9.6-4 between the two sinc terms at $\pm f_c$, as is appropriate for $f_c T >> 1$.

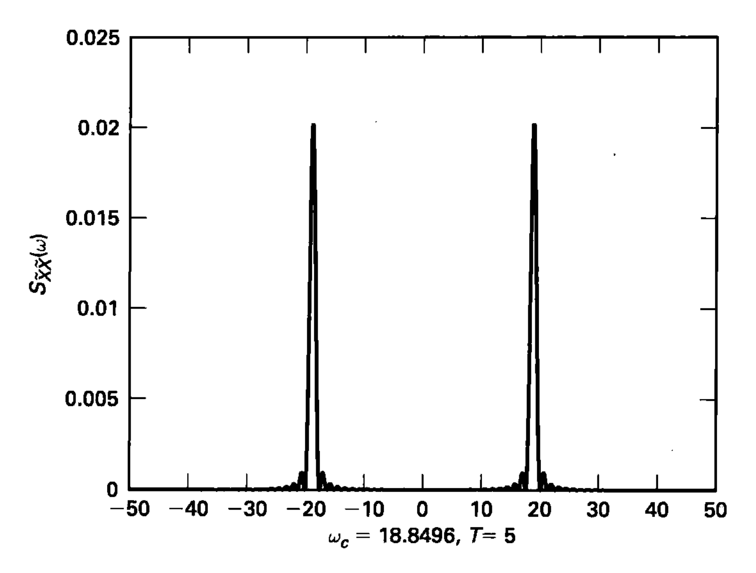


Figure 9.6-5 Power spectral density of PSK plotted for $f_c = 3$ and $T = 5$.

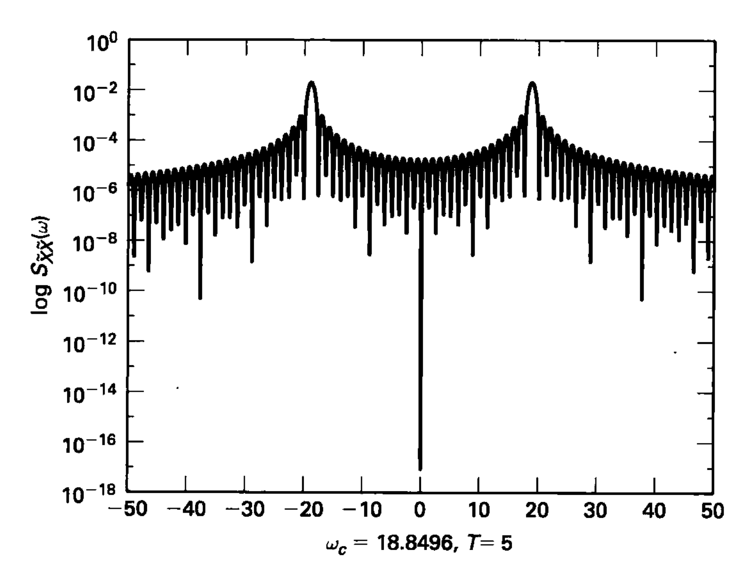


Figure 9.6-6 Log of power spectral density of PSK plotted for $f_c = 3$ and $T = 5$.

9.7 VECTOR PROCESSES AND STATE EQUATIONS

In this section we will generalize some of the results of Section 9.5 to the important class of vector random processes. This will lead into a brief discussion of state equations and vector Markov processes. Vector random processes occur in two-channel systems that are used in communications to model the in-phase and quadrature components of bandpass signals. Vector processes are also used extensively in control systems to model industrial processes with several inputs and outputs. Also, vector models are created artificially from high-order scalar models in order to employ the useful concept of *state* in both estimation and control theory.

Let $X_1(t)$ and $X_2(t)$ be two jointly stationary random processes that are input to the systems H_1 and H_2, respectively. Call the outputs Y_1 and Y_2, as shown in Figure 9.7-1.

From earlier discussions we know how to calculate $R_{X_1Y_1}$, $R_{X_2Y_2}$, $R_{Y_1Y_1}$, $R_{Y_2Y_2}$. We now look at how to calculate the correlations *across* the systems, that is, $R_{X_1Y_2}$, $R_{X_2Y_1}$, and $R_{Y_1Y_2}$. Given $R_{X_1X_2}$, we first calculate

$$\begin{aligned}
R_{X_1Y_2}(\tau) &= E[X_1(t+\tau)Y_2^*(t)] \\
&= \int_{-\infty}^{+\infty} E[X_1(t+\tau)X_2^*(t-\beta)]h_2^*(\beta)d\beta \\
&= \int_{-\infty}^{+\infty} R_{X_1X_2}(\tau+\beta)h_2^*(\beta)d\beta \\
&= \int_{-\infty}^{+\infty} R_{X_1X_2}(\tau-\beta')h_2^*(-\beta')d\beta', \qquad (\beta'=-\beta),
\end{aligned}$$

so

$$R_{X_1Y_2}(\tau) = R_{X_1X_2}(\tau) * h_2^*(-\tau),$$

and by symmetry

$$R_{X_2Y_1}(\tau) = R_{X_2X_1}(\tau) * h_1^*(-\tau).$$

The cross-correlation at the outputs is

$$R_{Y_1Y_2}(\tau) = h_1(\tau) * R_{X_1X_2}(\tau) * h_2^*(-\tau).$$

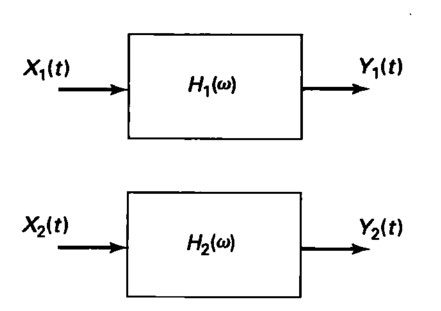


Figure 9.7-1 A generic (uncoupled) two-channel LSI system.

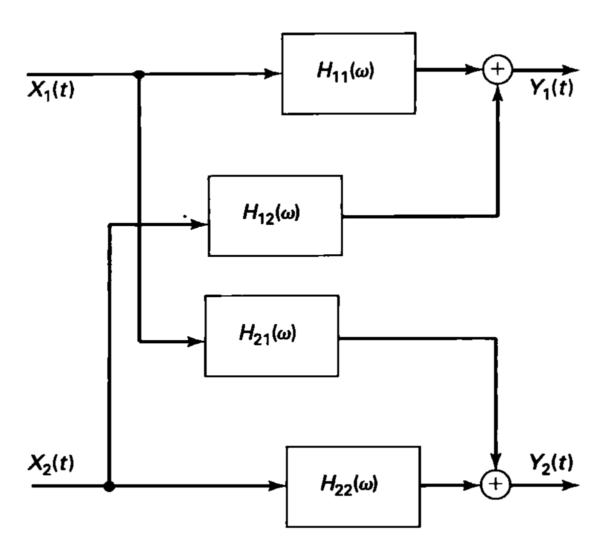


Figure 9.7-2 General two-channel LSI system.

Expressing these results in the spectral domain, we have

$$S_{X_1Y_2}(\omega) = S_{X_1X_2}(\omega)H_2^*(\omega)$$

and

$$S_{Y_1Y_2}(\omega) = H_1(\omega)H_2^*(\omega)S_{X_1X_2}(\omega).$$

In passing, we note the following important fact: If the *supports*† of the two system functions H_1 and H_2 do not overlap, then Y_1 and Y_2 are orthogonal random processes independent of any correlation in the input processes. We can generalize the above to a two-channel system with internal coupling as seen in Figure 9.7-2. Here two additional system functions have been added to cross-couple the inputs and outputs. They are denoted by H_{12} and H_{21}.

This case is best treated with vector notation; thus we define

$$\mathbf{X}(t) \triangleq [X_1(t), X_2(t)]^T, \qquad \mathbf{Y}(t) \triangleq [Y_1(t), Y_2(t)]^T,$$

and

$$\mathbf{h}(t) \triangleq \begin{bmatrix} h_{11}(t) & h_{12}(t) \\ h_{21}(t) & h_{22}(t) \end{bmatrix},$$

where $h_{ij}(t)$ is the impulse response of the subsystem with frequency response $H_{ij}(\omega)$. We then have

$$\mathbf{Y}(t) = \mathbf{h}(t) * \mathbf{X}(t), \tag{9.7-1}$$

†We recall that the support of a function g is defined as

$$\text{supp}(g) \triangleq \{x | g(x) \neq 0\}.$$

where the *vector convolution* is defined by

$$(\mathbf{h}(t) * \mathbf{X}(t))_i \triangleq \sum_{j=1}^{N} h_{ij}(t) * X_j(t).$$

If we define the following relevant input and output correlation matrices

$$\mathbf{R_{XX}}(\tau) \triangleq \begin{bmatrix} R_{X_1X_1}(\tau) & R_{X_1X_2}(\tau) \\ R_{X_2X_1}(\tau) & R_{X_2X_2}(\tau) \end{bmatrix} \tag{9.7-2}$$

$$\mathbf{R_{YY}}(\tau) \triangleq \begin{bmatrix} R_{Y_1Y_1}(\tau) & R_{Y_1Y_2}(\tau) \\ R_{Y_2Y_1}(\tau) & R_{Y_2Y_2}(\tau) \end{bmatrix}, \tag{9.7-3}$$

one can show that (Problem 9.44)

$$\mathbf{R_{YY}}(\tau) = \mathbf{h}(\tau) * \mathbf{R_{XX}}(\tau) * \mathbf{h}^{\dagger}(-\tau), \tag{9.7-4}$$

where the $\dagger$ indicates the Hermitian (or conjugate) transpose.

Taking the matrix Fourier transformation, we obtain

$$\mathbf{S_{YY}}(\omega) = \mathbf{H}(\omega)\mathbf{S_{XX}}(\omega)\mathbf{H}^{\dagger}(\omega) \tag{9.7-5}$$

with

$$\mathbf{H}(\omega) = FT\{\mathbf{h}(t)\},$$

and

$$\mathbf{S}(\omega) = FT\{\mathbf{R}(\tau)\},$$

where this notation is meant to imply an *element-by-element Fourier transform.* This multichannel generalization clearly extends to the M input and N output case by just enlarging the matrix dimensions accordingly.

State Equations

As shown in Problem 9.43, it is possible to rewrite an Nth-order LCCDE in the form of a first-order vector differential equation where the dimension of the output vector is equal to N,

$$\dot{\mathbf{Y}}(t) = \mathbf{AY}(t) + \mathbf{BX}(t), \qquad -\infty < t < +\infty. \tag{9.7-6}$$

This is just a multichannel system as seen in Equation 9.7-1 and can be interpreted as a set of N coupled first-order LCCDEs. We can take the vector Fourier transform and calculate the system function

$$\mathbf{H}(\omega) = (j\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \tag{9.7-7}$$

to specify this LSI operation in the frequency domain. Here $\mathbf{I}$ is the identity matrix. Alternately, we can express the operation in terms of a matrix convolution

$$\mathbf{Y}(t) = \mathbf{h}(t) * \mathbf{X}(t),$$

where we assume the multichannel system is stable; that is, all the impulse responses h_{ij} are BIBO stable. The solution proceeds much the same as in the scalar case for the first-order equation; in fact, it can be shown that

$$\mathbf{h}(t) = \exp(\mathbf{A}t)\,\mathbf{B}u(t). \tag{9.7-8}$$

The matrix exponential function $\exp(\mathbf{A}t)$ was encountered earlier in this chapter in the solution of the probability vector for a continuous-time Markov chain. This function is widely used in linear system theory, where its properties have been studied extensively [9–3].

If we compute the cross-correlation matrices in the WSS case, we obtain

$$\mathbf{R_{YX}}(\tau) = \exp(\mathbf{A}\tau)\,\mathbf{B}u(\tau) * \mathbf{R_{XX}}(\tau)$$

and

$$\mathbf{R_{XY}}(\tau) = \mathbf{R_{XX}}(\tau) * \mathbf{B}^{\dagger}\exp(-\mathbf{A}^{\dagger}\tau)u(-\tau),$$

with output correlation matrix, as before,

$$\mathbf{R_{YY}}(\tau) = \mathbf{h}(\tau) * \mathbf{R_{XX}}(\tau) * \mathbf{h}^{\dagger}(-\tau). \tag{9.7-9}$$

Upon vector Fourier transformation, this becomes

$$\mathbf{S_{YY}}(\omega) = (j\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{S_{XX}}(\omega)\mathbf{B}^{\dagger}(-j\omega\mathbf{I} - \mathbf{A}^{\dagger})^{-1}. \tag{9.7-10}$$

If $\mathbf{R_{XX}}(\tau) = \mathbf{Q}\delta(\tau)$, then since the system $\mathbf{H}$ is assumed causal, that is, $\mathbf{h}(t) = \mathbf{0}$ for $t < 0$, we have that the cross-correlation matrix $\mathbf{R}_{\mathbf{YX}}(\tau) = \mathbf{0}$ for $\tau < 0$; that is, $E[\mathbf{Y}(t+\tau)\mathbf{X}^{\dagger}(t)] = \mathbf{0}$ for $\tau < 0$. In words we say that $\mathbf{Y}(t+\tau)$ is orthogonal to $\mathbf{X}(t)$ for $\tau < 0$. Thus, the past of $\mathbf{Y}(t)$ is orthogonal to the present and future of $\mathbf{X}(t)$. If we additionally assume that the input process $\mathbf{X}(t)$ is a Gaussian process, then the uncorrelatedness condition becomes an independence condition. Under the Gaussian assumption then, the output $\mathbf{Y}(t)$ is independent of the present and future of $\mathbf{X}(t)$. A similar result was noted earlier in the scalar-valued case. We can use this result to show that the solution to a first-order vector LCCDE is a vector Markov random process with the following definition.

Definition 9.7-1 (*vector Markov*) A random process $\mathbf{Y}(t)$ is vector Markov if for all $n > 0$ and for all $t_n > t_{n-1} > \ldots > t_1$, and for all values $\mathbf{y}(t_{n-1}), \ldots, \mathbf{y}(t_1)$, we have

$$P[\mathbf{Y}(t_n) \le \mathbf{y}_n | \mathbf{y}(t_{n-1}), \ldots, \mathbf{y}(t_1)] = P[\mathbf{Y}(t_n) \le \mathbf{y}_n | \mathbf{y}(t_{n-1})]$$

for all values of the real vector $\mathbf{y}_n$. Here $\mathbf{A} \le \mathbf{a}$ means

$$(A_n \le a_n, A_{n-1} \le a_{n-1}, \ldots, A_1 \le a_1). \quad \blacksquare$$

Before discussing vector differential equations we briefly recall a result for *deterministic* vector LCCDEs. The first-order vector equation,

$$\dot{\mathbf{y}}(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{B}\mathbf{x}(t), \qquad t \ge t_0,$$

subject to the initial condition $\mathbf{y}(t_0)$, can be shown to have solution, employing the matrix exponential

$$\mathbf{y}(t) = \exp[\mathbf{A}(t - t_0)]\mathbf{y}(t_0) + \int_{t_0}^{t} \mathbf{h}(t - v)\mathbf{x}(v)dv, \qquad t \geq t_0,$$

thus generalizing the scalar case. This deterministic solution can be found in any graduate text on linear systems theory, for example, in [9–3]. The first term is called the *zero-input solution* and the second term is called the *zero-state* (or *driven) solution* analogously to the solution for scalar LCCDEs.

We can extend this theory to the stochastic case by considering the differential Equation 9.7-6 over the semi-infinite domain $t_0 \leq t < \infty$ and replacing the above deterministic solution with the following stochastic solution, expressed with the help of an integral:

$$\mathbf{Y}(t) = \exp[\mathbf{A}(t - t_0)]\mathbf{Y}(t_0) + \int_{t_0}^{t} \mathbf{h}(t - v)\mathbf{X}(v)dv. \tag{9.7-11}$$

If the LCCDE is BIBO stable, that is, the real parts of the eigenvalues of $\mathbf{A}$ are all negative, in the limit as $t_0 \to -\infty$, we get the solution for all time, that is $t_0 = -\infty$,

$$\mathbf{Y}(t) = \int_{-\infty}^{t} \mathbf{h}(t - v)\mathbf{X}(v)dv = \mathbf{h}(t) * \mathbf{X}(t), \tag{9.7-12}$$

which is the same as already derived for the stationary infinite time-interval case. In effect, we use the stability of the system to conclude that the resulting zero-input part of the solution must be zero at any finite time.

The following theorem shows a method to generate a vector Gauss–Markov random process using the above approach. The input is now a white Gaussian vector process $\mathbf{W}(t)$ and the output vector Markov process is denoted by $\mathbf{X}(t)$.

Theorem 9.7-1 Let the input to the state equation

$$\dot{\mathbf{X}}(t) = \mathbf{A}\mathbf{X}(t) + \mathbf{B}\mathbf{W}(t)$$

be the white Gaussian process $\mathbf{W}(t)$. Then the output $\mathbf{X}(t)$ is a vector Gauss–Markov random process.

Proof We write the solution at t_n in terms of the solution at an earlier time t_{n-1} as

$$\mathbf{X}(t_n) = \exp[\mathbf{A}(t_n - t_{n-1})]\mathbf{X}(t_{n-1}) + \int_{t_{n-1}}^{t_n} \mathbf{h}(t_n - v)\mathbf{W}(v)dv.$$

Then we write the integral term as $\mathbf{I}(t_n)$ and note that it is independent of $\mathbf{X}(t_{n-1})$. Thus we can deduce that

$$\begin{aligned} &P[\mathbf{X}(t_n) \leq \mathbf{x}_n|\mathbf{x}(t_{n-1}), \ldots, \mathbf{x}(t_1)] \\ &\quad = P[\mathbf{I}(t_n) \leq \mathbf{x}_n - e^{\mathbf{A}(t_n - t_{n-1})}\mathbf{x}(t_{n-1})|\mathbf{x}(t_{n-1}), \ldots, \mathbf{x}(t_1)] \\ &\quad = P[\mathbf{I}(t_n) \leq \mathbf{x}_n - e^{\mathbf{A}(t_n - t_{n-1})}\mathbf{x}(t_{n-1})|\mathbf{x}(t_{n-1})] \end{aligned}$$

and hence that $\mathbf{X}(t)$ is a vector Markov process. ■

If in Theorem 9.7-1 we did not have the Gaussian condition on the input $\mathbf{W}(t)$ but just the white noise condition, then we could not conclude that the output was Markov. This is because we would not have the independence condition required in the proof but only the weaker uncorrelatedness condition. On the other hand, if we relax the Gaussian condition but require that the input $\mathbf{W}(t)$ be an independent random process, then the process $\mathbf{X(t)}$ would still be Markov, but not Gauss-Markov. We use $\mathbf{X}$ for the process in this theorem rather than $\mathbf{Y}$ to highlight the fact that LCCDEs are often used to model input processes too.

SUMMARY

In this chapter we introduced the concept of the random process, an ensemble of functions of a continuous parameter. The parameter is most often time but can be position or another continuous variable. Most topics in this chapter generalize to two- and three-dimensional parameters. Many modern applications, in fact, require a two-dimensional parameter, for example, the intensity function $i(t_1, t_2)$ of an image. Such random functions are called *random fields* and can be analyzed using extensions of the methods of this chapter. Random fields are discussed in Chapter 7 of [9-5] and in [9-8] among many other places.

We introduced a number of important processes: asynchronous binary signaling; the Poisson counting process; the random telegraph signal; phase-shift keying, which is basic to digital communications; the Wiener process, our first example of a Gaussian random process and a basic building block process in nonlinear filter theory; and the Markov process, which is widely used for its efficiency and tractability and is the signal model in the widely employed Kalman–Bucy filter of Chapter 11.

We considered the effect of linear systems on the second-order properties of random processes. We specialized our results to the useful subcategory of stationary and WSS processes and introduced the power spectral density and the corresponding analysis for LSI systems. We also briefly considered the classes of wide-sense periodic and cyclostationary processes and introduced random vector processes and systems and extended the Markov model to them.

PROBLEMS

(*Starred problems are more advanced and may require more work and/or additional reading.)

9.1 Let $X[n]$ be a real-valued stationary random sequence with mean $E\{X[n]\} = \mu_X$ and autocorrelation function $E\{X[n+m]X[n]\} = R_{XX}[m]$. If $X[n]$ is the input to a D/A converter, the continuous-time output can be idealized as the *analog* random process $X_a(t)$ with

$$X_a(t) \triangleq X[n] \quad \text{for } n \le t < n+1, \quad \text{for all } n,$$

as shown in Figure P9.1.

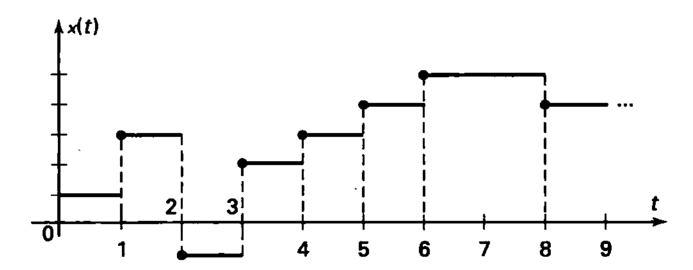


Figure P9.1 Typical output of sample-hold D/A converter.

(a) Find the mean $E[X_a(t)] = \mu_a(t)$ as a function of μ_X.
(b) Find the correlation $E[X_a(t_1)X_a(t_2)] = R_{X_aX_a}(t_1, t_2)$ in terms of $R_{XX}[m]$.

9.2 Consider a WSS random sequence $X[n]$ with mean function μ_X, a constant, and correlation function $R_{XX}[m]$. Form a random process as

$$X(t) \triangleq \sum_{n=-\infty}^{+\infty} X[n] \frac{\sin \pi(t-nT)/T}{\pi(t-nT)/T}, \qquad -\infty < t < +\infty.$$

In what follows, we assume the infinite sums converge and so, do not worry about stochastic convergence issues.

(a) Find $\mu_X(t)$ in terms of μ_X. Simplify your answer as much as possible.
(b) Find $R_{XX}(t_1, t_2)$ in terms of $R_{XX}[m]$. Is $X(t)$ WSS?

Hint: The sampling theorem from *Linear Systems Theory* states that any bandlimited deterministic function $g(t)$ can be recovered exactly from its evenly spaced samples, that is,

$$g(t) = \sum_{n=-\infty}^{+\infty} g(nT) \frac{\sin \pi(t-nT)/T}{\pi(t-nT)/T},$$

when the radian bandwidth of the function $g(t)$ is π/T or less.

9.3 Consider the random process $Y(t) = (-1)^{X(t)}$, where $X(t)$ is a Poisson process with rate λ. Thus, $Y(t)$ starts at $Y(0) = 1$ and switches back and forth from $+1$ to -1 at random Poisson times T_i

(a) Find the mean of $Y(t)$
(b) Find the autocorrelation function of $Y(t)$
(c) If $Z(t) = AY(t)$

where A is a random variable independent of $Y(t)$ and takes on the values ± 1 with equal probability, show that $Z(t)$ is *WSS* and find the power spectral density of $Z(t)$.

9.4 The output $Y(t)$ of a tapped delay-line filter shown in Figure P9.4, with input $X(t)$ and N taps, is given by

$$Y(t) = \sum_{n=0}^{N-1} A_n X(t - nT).$$

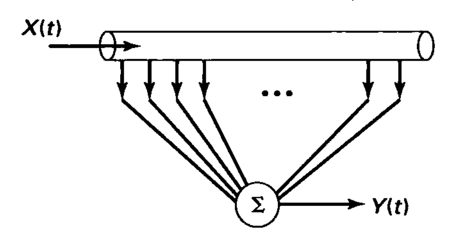


Figure P9.4 Tapped delay-line filter.

The input $X(t)$ is a stationary Gaussian random process with zero mean and autocorrelation function $R_{XX}(\tau)$ having the property that $R_{XX}(nT) = 0$ for every integer $n \neq 0$. The tap gains $A_n, n = 0, 1, \ldots, N-1$, are zero-mean, uncorrelated Gaussian random variables with common variance σ_A^2. Every tap gain is independent of the input process $X(t)$.

(a) Find the autocorrelation function of $Y(t)$.
(b) For a given value of t, find the characteristic function of $Y(t)$. Justify your steps.
(c) For fixed t, what is the asymptotic pdf of $\frac{1}{\sqrt{N}}Y(t)$, asymptotic as $N \to \infty$? Explain.
(d) Suppose now that the number of taps N is a Poisson random variable with mean $\lambda(> 0)$. Find the answers to parts (a) and (b) now.

(*Note*: You may need to use the following: $e^{-x} \approx \frac{1}{1+x}$ for $|x| << 1$, and $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.)

9.5 Let $N(t)$ be a Poisson random process defined on $0 \leq t < \infty$ with $N(0) = 0$ and mean arrival rate $\lambda > 0$.

(a) Find the joint probability $P[N(t_1) = n_1, N(t_2) = n_2]$ for $t_2 > t_1$.
(b) Find an expression for the Kth order joint PMF,

$$P_N(n_1, \ldots, n_K; t_1, \ldots, t_K),$$

with $0 \leq t_1 < t_2 < \ldots < t_K < \infty$. Be careful to consider the relative values of $n_1, \ldots, n_K$.

***9.6** The *nonuniform Poisson counting process* $N(t)$ is defined for $t \geq 0$ as follows:

(a) $N(0) = 0$.
(b) $N(t)$ has independent increments.
(c) For all $t_2 \geq t_1$,

$$P[N(t_2) - N(t_1) = n] = \frac{\left[\int_{t_1}^{t_2} \lambda(v)dv\right]^n}{n!} \exp\left(-\int_{t_1}^{t_2} \lambda(v)dv\right), \quad \text{for } n \geq 0.$$

The function $\lambda(t)$ is called the *intensity function* and is everywhere nonnegative, that is, $\lambda(t) \geq 0$ for all t.

(a) Find the mean function $\mu_N(t)$ of the nonuniform Poisson process.
(b) Find the correlation function $R_{NN}(t_1, t_2)$ of $N(t)$. Define a *warping* of the time axis as follows:

$$\tau(t) \triangleq \int_0^t \lambda(v)dv.$$

Now $\tau(t)$ is monotonic increasing if $\lambda(v) > 0$ for all v, so we can then define the inverse mapping $t(\tau)$ as shown in Figure P9.6.

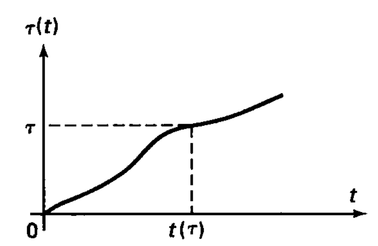


Figure P9.6 Plot of τ versus t.

(c) Assume $\lambda(t) > 0$ for all t and define the counting process,

$$N_u(\tau) \triangleq N(t(\tau)).$$

Show that $N_u(\tau)$ is a uniform Poisson counting process with rate $\lambda = 1$; that is, show for $\tau \geq 0$

(1) $N_u(0) = 0$.
(2) $N_u(\tau)$ has independent increments.
(3) For all $\tau_2 \geq \tau_1$,

$$P[N_u(\tau_2) - N_u(\tau_1) = n] = \frac{(\tau_2 - \tau_1)^n}{n!} e^{-(\tau_2 - \tau_1)} \qquad n \geq 0.$$

9.7 A nonuniform Poisson process $N(t)$ has intensity function (mean arrival rate)

$$\lambda(t) = 1 + 2t,$$

for $t \geq 0$. Initially $N(0) = 0$.

(a) Find the mean function $\mu_N(t)$.
(b) Find the correlation function $R_{NN}(t_1, t_2)$.
(c) Find an expression for the probability that $N(t) \geq t$, that is, find $P[N(t) \geq t]$ for any $t > 0$.
(d) Give an approximate answer for (c) in terms of the error function $\text{erf}(x)$.

***9.8** This problem concerns the construction of the Poisson counting process as given in Section 9.2.

(a) Show the density for the nth arrival time $T[n]$ is

$$f_T(t; n) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} u(t), \qquad n > 0.$$

In the derivation of the property that the increments of a Poisson process are Poisson distributed, that is,

$$P[X(t_a) - X(t_b) = n] = \frac{[\lambda(t_a - t_b)]^n}{n!} e^{-\lambda(t_a - t_b)} u[n], \qquad t_a > t_b,$$

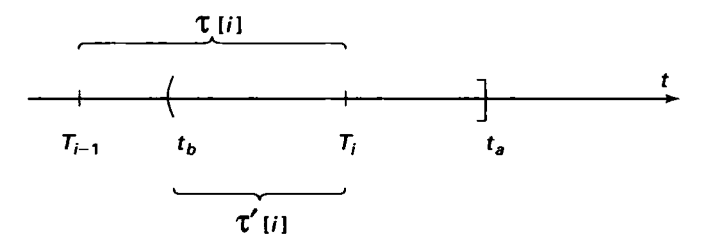


Figure P9.8 Illustrative example of relation of arrival times to arbitrary observation interval.

we implicitly use the fact that the first interarrival time in $(t_b, t_a]$ is exponentially distributed. Actually, this fact is not clear as the interarrival time in question is only partially in the interval $(t_b, t_a]$. A pictorial diagram is shown in Figure P9.8. Define $\tau'[i] \triangleq T[i] - t_b$ as the *partial interarrival time.* We note $\tau'[i] = \tau[i] - T$, where the random variable $T \triangleq t_b - T[i-1]$ and $\tau[i]$ denotes the (full) interarrival time.

(b) Fix the random variable $T = t$ and find the CDF

$$F_{\tau'[i]}(\tau'|T = t) = P\{\tau[i] \leq \tau' + t | \tau[i] \geq t\}.$$

(c) Modify the result of part (b) to account for the fact that T is a random variable, and find the unconditional CDF of τ'. (*Hint*: This part does not involve a lot of calculations.)

Because of the preceding properties, the exponential distribution is called *memoryless.* It is the only continuous distribution with this property.

9.9 Let $N(t)$ be a counting process on $[0, \infty)$ whose average rate $\lambda(t)$ depends on another positive random process $S(t)$, specifically $\lambda(t) = S(t)$. We assume that $N(t)$ given $\{S(t) \text{ on } [0, \infty)\}$ is a *nonuniform* Poisson process. We know $\mu_S(t) = \mu_0 > 0$ and also know $K_{SS}(t_1, t_2)$.

(a) Find $\mu_N(t)$ for $t \geq 0$ in terms of μ_0.
(b) Find $\sigma_N^2(t)$ for $t \geq 0$ in terms of $K_{SS}(t_1, t_2)$.

9.10 Let the random process $K(t)$ (not a covariance!) depend on a uniform Poisson process $N(t)$, with mean arrival rate $\lambda > 0$, as follows: Starting at $t = 0$, both $N(t) = 0$ and $K(t) = 0$. When an arrival occurs in $N(t)$, an independent Bernoulli trial takes place with probability of success p, where $0 < p < 1$. On success, $K(t)$ is incremented by 1, otherwise $K(t)$ is left unchanged. This arrangement is shown in Figure P9.11. Find the first-order PMF of the discrete-valued random process $K(t)$ at time t, that is, $P_K(k; t)$, for $t \geq 0$.

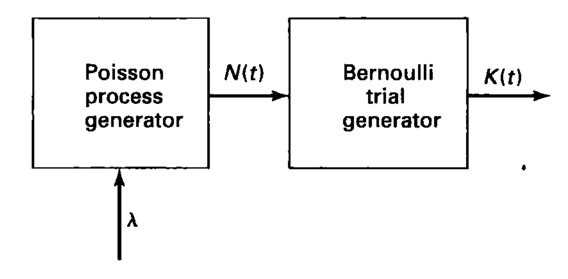


Figure P9.11 Poisson-modulated Bernoulli trial process.

9.11 Let the scan-line of an image be described by the spatial random process $S(x)$, which models the ideal gray level at the point x. Let us transmit each point independently with an optical channel by modulating the intensity of a photon source:

$$\lambda(t, x) = S(x) + \lambda_0, \qquad 0 \leq t \leq T.$$

In this way we create a family of random processes, indexed by the continuous parameter x,

$$\{N(t, x)\}.$$

For each x, $N(t, x)$ given $S(x)$ is a uniform Poisson process. At the end of the observation interval, we store $N(x) \triangleq N(T, x)$ and inquire about the statistics of this spatial process.

To summarize, $N(x)$ is an integer-valued *spatial* random process that depends on the value of another random process $S(x)$, called the signal process. The spatial random process $S(x)$ is stationary with zero mean and covariance function

$$K_{SS}(x) = \sigma_S^2 \exp(-\alpha|x|),$$

where $\alpha > 0$. The conditional distribution of $N(x)$, given $S(x) = s(x)$, is Poisson with mean $\lambda(x) = (s(x) + \lambda_0)T$, where λ_0 is a positive constant; that is,

$$P[N(x) = n | S(x) = s(x)] = \frac{\lambda^n(x)}{n!} e^{-\lambda(x)} u[n].$$

The random variables $N(x)$ are conditionally independent from point to point.

(a) Find the (unconditional) mean and variance

$$\mu_N(x) = E[N(x)] \quad \text{and} \quad E\left[(N(x) - \mu_N(x))^2\right].$$

(*Hint*: First find the conditional mean and conditional mean square.)

(b) Find $R_{NN}(x_1, x_2) \triangleq E[N(x_1)N(x_2)]$.

9.12 Let $X(t)$ be a random telegraph signal (RTS) defined on $t \geq 0$. Fix $X(0) = +1$. The RTS uses a Poisson random arrival time sequence $T[n]$ to switch the value of $X(t)$ between ± 1. Take the average arrival rate as $\lambda (> 0)$. Thus we have

$$X(t) \triangleq \begin{cases} 1, & 0 \leq t < T[1] \\ -1, & T[1] \leq t < T[2] \\ +1, & T[2] \leq t < T[3] \\ \cdots, & \cdots \end{cases}.$$

(a) Argue that $X(t)$ is a Markov process and draw and label the state-transition diagram.

(b) Find the steady-state probability that $X(t) = +1$, that is, $P_X(1; \infty)$, in terms of the rate parameter λ.

(c) Write the differential equations for the state probabilities $P_X(1;t)$ and $P_X(-1;t)$.

***9.13** A uniform Poisson process $N(t)$ with rate $\lambda (> 0)$ is an infinite-state Markov chain with the state-transition diagram in Figure P9.13a. Here the state labels are the values of the process (chain) $N(t)$ between the transitions. Also the independent interarrival times $\tau[n]$ are exponentially distributed with parameter λ.

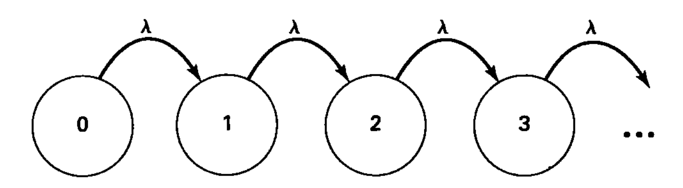


Figure P9.13a Poisson process represented as Markov chain.

We make the following modifications to the above scenario. Replace the independent interarrival times $\tau[n]$ by an arbitrary nonnegative, stationary, and independent random sequence, still denoted $\tau[n]$, resulting in the generalization called a *renewal process* in the literature. See Figure P9.13b.

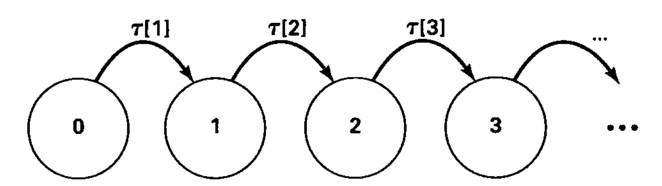


Figure P9.13b More general (renewal) process chain.

(a) Show that the PMF $P_N(n;t) = P[N(t) = n]$ of a renewal process is given, in terms of the CDF of the arrival times $F_T(t;n)$, as

$$P_N(n;t) = F_T(t;n) - F_T(t;n+1), \qquad \text{when } n \geq 1,$$

where the arrival time $T[n] = \sum_{k=1}^{n} \tau[k]$ and $F_T(t;n)$ is the corresponding CDF of the arrival time $T[n]$.

(b) Let $\tau[n]$ be $U[0,1]$, that is, uniformly distributed over $[0,1]$, and find $P_N(n;t)$ for $n = 0, 1$, and 2, for this specific renewal process.

(c) Find the characteristic function of the renewal process of part (b).

(d) Find an approximate expression for the CDF $F_T(t;n)$ of the renewal process in part (b), that is good for large n, and not too far from the $T[n]$ mean value. (*Hint*: For small x we have the trigonometric series approximation $\sin x \approx x - x^3/3!$)

9.14 If $X(t)$ with $X(0) = 0$ and $\mu = 0$ is a Wiener process, show that $Y(t) = \sigma X(t/\sigma^2)$ is also a Wiener process. Find its covariance function.

9.15 Let $W_1(t)$ and $W_2(t)$ be two Wiener processes, independent of one another, both defined on $t \geq 0$, with variance parameters α_1 and α_2, respectively. Let the process $X(t)$ be defined as their algebraic difference, that is, $X(t) \triangleq W_1(t) - W_2(t)$.

(a) What is $R_{XX}(t_1, t_2)$ for $t_1, t_2 \geq 0$?

(b) What is the pdf $f_X(x;t)$ for $t \geq 0$?

***9.16** If the $2n$ random variables A_r and B_r are uncorrelated with zero mean and $E(A_r^2) = E(B_r^2) = \sigma_r^2$, show that the random process

$$X(t) = \sum_{r=1}^{n} (A_r \cos w_r t + B_r \sin w_r t)$$

is wide-sense stationary. What are the mean and autocorrelation of $X(t)$?

***9.17** Let $W(t)$ be a standard Wiener process, that is, $\alpha = 1$, and define

$$X(t) \triangleq W^2(t) \quad \text{for} \quad t \geq 0.$$

(a) Find the probability density $f_X(x;t)$.

(b) Find the conditional probability density $f_X(x_2|x_1;t_2,t_1), \qquad t_2 > t_1$.

(c) Is $X(t)$ Markov? Why?

(d) Does $X(t)$ have independent increments? Justify.

9.18 Let $X(t)$ be a Markov random process on $[0, \infty)$ with initial density $f_X(x;0) = \delta(x-1)$ and conditional pdf

$$f_X(x_2|x_1;t_2,t_1) = \frac{1}{\sqrt{2\pi(t_2-t_1)}} \exp\left(-\frac{1}{2}\frac{(x_2-x_1)^2}{t_2-t_1}\right), \qquad \text{for all } t_2 > t_1.$$

(a) Find $f_X(x;t)$ for all t.
(b) Repeat part (a) for $f_X(x;0) \sim N(0,1)$.

9.19 Consider the three-state Markov chain $N(t)$ with the state-transition diagram shown in Figure P9.19. Here the state labels are the actual outputs, eg. $N(t) = 3$, while the chain is in state 3. The state transitions are governed by jointly independent, exponentially distributed interarrival times, with average rates as indicated on the branches.

(a) *Given* that we start in state 2 at time $t = 0$, what is the probability (conditional probability) that we *remain* in this state until time t, for some arbitrary $t > 0$? (*Hint*: There are two ways to leave state 2. So you will leave at the lesser of the two independent exponential random variables with rates μ_2 and λ_2.)
(b) Write the differential equations for the probability of being in state i at time $t \geq 0$, denoting them as $p_i(t)$, $i = 1, 2, 3$. [*Hint*: First write $p_i(t + \delta t)$ in terms of the $p_i(t)$, $i = 1, 2, 3$, only keeping terms up to order $O(\delta t)$.]
(c) Find the steady-state solution for $p_i(t)$ for $i = 1, 2, 3$, that is, $p_i(\infty)$.

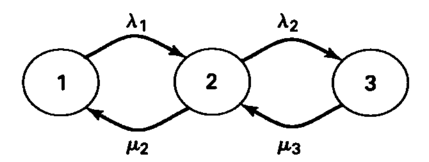


Figure P9.19 A three-state continuous-time Markov chain.

9.20 Let a certain wireless communication binary channel be in a *good* state or *bad* state, described by the continuous-time Markov chain with transition rates as shown in Figure P9.20. Here we are given that the exponentially distributed state transitions have rates $\lambda_1 = 1$ and $\lambda_2 = 9$. The value of ϵ for each state is given in part (b) below.

(a) Find the steady-state probability that the channel is in *good* state. Label $P\{X(t) = \text{ good }\} = p_G$, and $P\{X(t) = \text{ bad }\} = p_B$. (*Hint*: Assume the steady state exists and then write p_G at time t in terms of the two possibilities at time $t - \delta$, keeping only terms to first order in δ, taken as very small.)
(b) Assume that in the good state, there are no errors on the binary channel, but in the bad state the probability of error is $\epsilon = 0.01$ Find the average error probability on the channel. (Assume that the channel does not change state during the transmission of each single bit.)

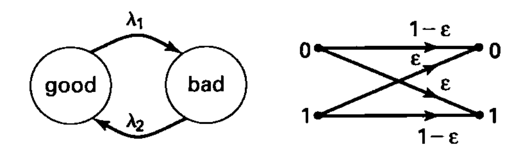


Figure P9.20 Model of two-state wireless communication channel.

9.21 This problem concerns the Chapman–Kolmogorov equation (cf. Equation 9.2-22) for a continuous-amplitude Markov random process $X(t)$,

$$f_X(x(t_3)|x(t_1)) = \int_{-\infty}^{+\infty} f_X(x(t_3)|x(t_2))f_X(x(t_2)|x(t_1))\,dx(t_2),$$

for the conditional pdf at three increasing observation times $t_3 > t_2 > t_1 > 0$. You will show that the pdf of the Wiener process with covariance function $K_{XX}(t,s) = \alpha \min(t,s)$, $\alpha > 0$, solves the above equation.

(a) Write the first-order pdf $f_X(x(t))$ of this Wiener process for $t > 0$.
(b) Write the first-order conditional pdf $f_X(x(t)|x(s))$, $t > s > 0$.
(c) Referring back to the Chapman–Kolmogorov equation, set $t_3 - t_2 = t_2 - t_1 = \delta$ and use x_3, x_2, and x_1 to denote the values taken on. Then verify that your conditional pdf from part (b) satisfies the resulting equation

$$f_X(x_3|x_1) = \int_{-\infty}^{+\infty} f_X(x_3|x_2)f_X(x_2|x_1)\,dx_2.$$

9.22 Is the random process $X'(t)$ of Example 9.3-2 stationary? Why?

9.23 Let A and B be i.i.d. random variables with mean 0, variance σ^2, and third moment $m_3 \triangleq E[A^3] = E[B^3] \neq 0$. Consider the random process

$$X(t) = A\cos(2\pi ft) + B\sin(2\pi ft), \qquad -\infty < t < +\infty,$$

where f is a given frequency.

(a) Show that the random process $X(t)$ is WSS.
(b) Show that $X(t)$ is not strictly stationary.

9.24 Verify whether the sine-wave process $\{X(t)\}$ where $X(t) = Y$ cos wt, where w is a constant and Y is uniformly distributed in (0,1), is a strict sense stationary process.

9.25 If $X(t) = \mu + N(t)$, where $E[X(t)] = \mu, N(t)$ is a white noise with autocovariance function $K(t_1,t_2) = \phi(t_1)\delta(t_1 - t_2)$ where $Q(t)$ is a bounded function of t and Q is the unit impulse function, prove that $\{X(t)\}$ is a mean-ergodic process.

9.26 If $X(t)$ is a wide-sense stationary process with autocorrelation function $R_{XX}(\iota) = A_e^{-\alpha|\iota|}$, determine the second-order moment of the random variable $X(8) - X(5)$.

9.27 The power spectrum of a wide-sense stationary process $\{X(t)\}$ is given by $S_{XX}(w) = 1/(1+w^2)^2$. Find its autocorrelation function $R_{XX}(\iota)$ and average power.

9.28 Consider the LSI system shown in Figure P9.28, whose input is the zero-mean random process $W(t)$ and whose output is the random process $X(t)$. The frequency response of the system is $H(\omega)$. Given $K_{WW}(\tau) = \delta(\tau)$, find $H(\omega)$ in terms of the cross-covariance $K_{XW}(\tau)$ or its Fourier transform.

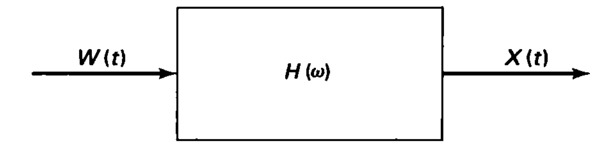


Figure P9.28 LSI system with white noise input.

9.29 If the input $x(t)$ and $y(t)$ are connected by the differential equation $T\frac{dy(t)}{dt} + y(t) = x(t)$, prove that they can be related by means of a convolution type integral. Assume that $x(t)$ and $y(t)$ are zero for $t \leq 0$.

9.30 Consider the first-order stochastic differential equation

$$\frac{dX(t)}{dt} + X(t) = W(t)$$

driven by the zero-mean white noise $W(t)$ with correlation function $R_{WW}(t,s) = \delta(t-s)$.

(a) If this differential equation is valid for all time, $-\infty < t < +\infty$, find the psd of the resulting wide-sense stationary process $X(t)$.

(b) Using residue theory (or any other method), find the inverse Fourier transform of $S_{XX}(\omega)$, the autocorrelation function $R_{XX}(\tau)$, $-\infty < \tau < +\infty$.

(c) If the above differential equation is run only for $t > 0$, is it possible to choose an initial condition random variable $X(0)$ such that $X(t)$ is wide-sense stationary for all $t > 0$? If such a random variable exists, find its mean and variance. You may assume that the random variable $X(0)$ is orthogonal to $W(t)$ on $t \geq 0$; that is, $X(0) \perp W(t)$. [*Hint*: Express $X(t)$ for $t > 0$ in terms of the initial condition and a stochastic integral involving $W(t)$.]

9.31 If the random process $\{X(t)\}$ is defined as $X(t) = Y(t)\, Z(t)$ where $Y(t)$ and $Z(t)$ are independent wide-sense stationary processes, determine the power spectral density of $X(t)$.

9.32 Consider a wide-sense stationary process $X(t)$ with autocorrelation function $R_{XX}(\iota)$ and power spectral density function $S_{XX}(w)$. Let $\dot{X}(t)\dfrac{dX(t)}{dt}$. Show that

(a) $R_{X\dot{X}}(t) = \dfrac{d}{d\tau} R_{XX}(\tau)$

(b) $R_{\dot{X}\dot{X}}(t) = \dfrac{-d^2}{d\tau^2} R_{XX}(\tau)$

(c) $S_{\dot{X}\dot{X}}(w) = w^2 S_{XX}(w)$

9.33 If $X(t)$ is the input voltage to a circuit (system) and $Y(t)$ is the output voltage where $X(t)$ is a stationary random process with mean $\mu_X = 0$ and autocorrelation function $R_{XX}(\iota) = e^{-\alpha|\iota|}$and if the power transfer function is $H(w) = \dfrac{R}{R + jLw}$, find the mean μ_Y, the autocorrelation function $R_{YY}(\iota)$ and the power spectral density $S_{YY}(w)$ of $Y(t)$.

9.34 A WSS and zero-mean random process $Y(t)$ has sample functions consisting of successive rectangular pulses of random amplitude and duration as shown in Figure P9.34.

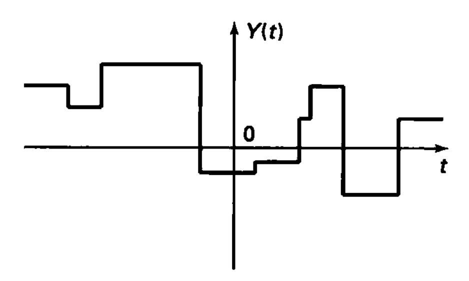


Figure P9.34 Random amplitude pulse train.

The pdf for the pulse width is

$$f_W(w) = \begin{cases} \lambda e^{-\lambda w}, & w \geq 0, \\ 0, & w < 0, \end{cases}$$

with $\lambda > 0$. The amplitude of each pulse is a random variable X (independent of W) with mean 0 and variance σ_X^2. Successive amplitudes and pulse widths are independent.

(a) Find the autocorrelation function $R_{YY}(\tau) = E[Y(t + \tau)Y(t)]$.

(b) Find the corresponding psd $S_{YY}(\omega)$.

[*Hint*: First find the *conditional* autocorrelation function $E[Y(t+\tau)Y(t)|W=w]$, where t is assumed to be at the start of a pulse (do this without loss of generality per WSS hypothesis for $Y(t)$).]

9.35 The power spectral density of a zero mean wide-sense stationary process $\{X(t)\}$ is given by

$$S_{XX}(w) = \begin{cases} 1, & |\mathrm{w}| < \mathrm{w}_o \\ 0, & \text{elsewhere} \end{cases}$$

Determine the autocorrelation function of $\{X(t)\}$ and show that $\{X(t)\}$ and $X\left(t+\frac{\tau}{W_0}\right)$ are uncorrelated.

***9.36** In this problem we consider using white noise as an approximation to a smoother process (cf. More on White Noise in Section 9.5), which is the input to a lowpass filter. The output process from the filter is then investigated to determine the error resulting from the white noise approximation. Let the stationary random process $X(t)$ have zero mean and autocovariance function

$$K_{XX}(\tau) = \frac{1}{2\tau_0}\exp(-|\tau|/\tau_0),$$

which can be written as $h(\tau) * h(-\tau)$ with $h(\tau) = \frac{1}{\tau_0}e^{-\tau/\tau_0}u(\tau)$.

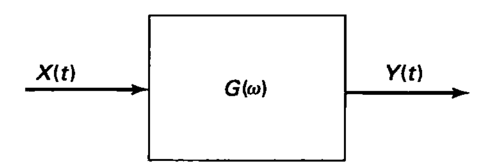


Figure P9.36a Approximation to white noise input to filter.

(a) Let $X(t)$ be input to the lowpass filter shown in Figure P9.36a, with output $Y(t)$. Find the output psd $S_Y(\omega)$, for

$$G(\omega) \triangleq \begin{cases} 1, & |\omega| \le \omega_0 \\ 0, & \text{else.} \end{cases}$$

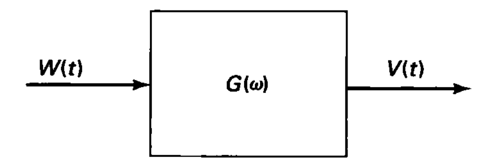


Figure P9.36b White noise input to filter.

(b) Alternatively we may, at least formally, excite the system directly with a standard white noise $W(t)$, with mean zero and $K_{WW}(\tau) = \delta(\tau)$. Call the output $V(t)$ as shown in Figure P9.36b. Find the output psd $S_{VV}(\omega)$.

(c) Show that for $|\omega_0\tau_0| << 1$, $S_{YY} \simeq S_{VV}$ and find an upper bound on the *power error*

$$|R_{VV}(0) - R_{YY}(0)|.$$

9.37 Consider the LSI system shown in Figure P9.37. Let $X(t)$ and $N(t)$ be WSS and mutually uncorrelated with power spectral densities $S_{XX}(\omega)$ and $S_{NN}(\omega)$ and zero means.

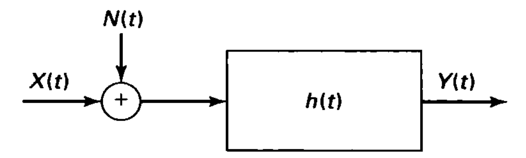


Figure P9.37

(a) Find the psd of the output $Y(t)$.
(b) Find the cross-power spectral density of X and Y, that is, find $S_{XY}(\omega)$ and $S_{YX}(\omega)$.
(c) Define the error $\xi(t) \triangleq Y(t) - X(t)$ and evaluate the psd of $\xi(t)$.
(d) Assume that $h(t) = a\delta(t)$ and choose the value of a which minimizes $E[\xi^2(t)] = R_{\xi\xi}(0)$.

9.38 Let $X(t)$ be a random process defined by

$$X(t) \triangleq N\cos(2\pi f_0 t + \Theta),$$

where f_0 is a known frequency and N and Θ are independent random variables. The CF for N is

$$\Phi_N(\omega) = E[e^{+j\omega N}] = \exp\{\lambda[e^{j\omega} - 1]\},$$

where λ is a given positive constant (i.e., N is a Poisson random variable). The random variable Θ is uniformly distributed on $[-\pi, +\pi]$.

(a) Determine the mean function $\mu_X(t)$.
(b) Determine the covariance function $K_{XX}(t, s)$.
(c) Is $X(t)$ WSS? Justify your answer.
(d) Is $X(t)$ stationary? Justify your answer.

9.39 Let $X(t)$ be an independent-increment random process defined on $t \geq 0$ with initial value $X(0) = X_0$, a random variable. Assume the following CFs exist: $E[e^{j\omega X_0}] \triangleq \Phi_{X_0}(\omega)$ and

$$E[e^{j\omega(X(t)-X_0(s))}] \triangleq \Phi_{X(t)-X_0(s)}(\omega) \quad \text{for} \quad t \geq s.$$

(a) On defining $E[e^{j\omega X(t)}] \triangleq \Phi_{X(t)}(\omega)$, show that

$$\Phi_{X(t)}(\omega) = \Phi_{X_0}(\omega)\Phi_{X(t)-X_0}(\omega).$$

(b) Show that for all $t_2 \geq t_1$, the joint characteristic function of $X(t_2)$ and $X(t_1)$ is given by

$$\Phi_{X(t_2),X(t_1)}(\omega_2, \omega_1) = \Phi_{X_0}(\omega_1 + \omega_2)\Phi_{X(t_1)-X_0}(\omega_1 + \omega_2)\Phi_{X(t_2)-X(t_1)}(\omega_2).$$

(c) Apply part (a) to Problem 9.18(b) by using Gaussian characteristic functions.

9.40 Given a random variable Y with characteristic function $\phi(w) = E[e^{iwy}]$ and a random process $X(t) = \cos(\lambda t + Y)$, where λ is a constant, show that $\{X(t)\}$ is stationary in the wide sense if $\phi(1) = \phi(2) = 0$.

9.41 Let $X(t)$ defined over $t \geq 0$ have independent increments with mean function $\mu_X(t) = \mu_0$ and covariance function

$$K_{XX}(t_1, t_2) = \sigma_X^2(\min(t_1, t_2)),$$

where $\sigma_X^2(t)$ is an increasing function, that is, $d\sigma_X^2(t)/dt > 0$ for all $t \geq 0$, called the *variance function.* Note that $\text{Var}[X(t)] = \sigma_X^2(t)$. Fix $T > 0$ and find the mean and covariance functions of $Y(t) \triangleq X(t) - X(T)$ for all $t \geq T$. (*Note*: For the covariance function take t_1 and $t_2 \geq T$.)

***9.42** Following Example 9.2-3, use MATLAB to compute a 1000-element sample function of the Wiener process $X(t)$ for $\alpha = 2$ and $T = 0.01$.

(a) Use the MATLAB routine `hist.m` to compute the histogram of $X(10)$ and compare it with the ideal Gaussian pdf.

(b) Estimate the mean of $X(10)$ using `mean.m` and the standard deviation using `std.m` and compare them to theoretical values. [*Hint*: Use `Wiener.m`† in a `for` loop to calculate 100 realizations of `x(1000)`. Then use `hist`. Question: Why can't you just use the last 100 elements of the vector `x` to approximately obtain the requested statistics?]

9.43 Let the WSS random process $X(t)$ be the input to the third-order differential equation

$$\frac{d^3Y}{dt^3} + a_2\frac{d^2Y}{dt^2} + a_1\frac{dY}{dt} + a_0Y(t) = X(t),$$

with WSS output random process $Y(t)$.

(a) Put this equation into the form of a first-order vector differential equation

$$\frac{d\mathbf{Y}}{dt} = \mathbf{AY}(t) + \mathbf{BX}(t),$$

by defining $\mathbf{Y}(t) \triangleq \begin{bmatrix} Y(t) \\ Y'(t) \\ Y''(t) \end{bmatrix}$ and $\mathbf{X}(t) \triangleq [X(t)]$ and evaluating the matrices **A** and **B**.

†`Wiener.m` is provided on this book's Web site.

(b) Find a first-order matrix-differential equation for $\mathbf{R}_{\mathbf{XY}}(\tau)$ with input $\mathbf{R}_{\mathbf{XX}}(\tau)$.
(c) Find a first-order matrix-differential equation for $\mathbf{R}_{\mathbf{YY}}(\tau)$ with input $\mathbf{R}_{\mathbf{XY}}(\tau)$.
(d) Using matrix Fourier transforms, show that the output psd matrix $\mathbf{S}_{\mathbf{YY}}$ is given as

$$\mathbf{S}_{\mathbf{YY}}(\omega) = (j\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{S}_{\mathbf{XX}}(\omega)\mathbf{B}^{\dagger}(-j\omega\mathbf{I} - \mathbf{A}^{\dagger})^{-1}.$$

9.44 Let $\mathbf{X}(t)$ be a WSS vector random process, which is input to the LSI system with impulse response matrix $\mathbf{h}(t)$.

(a) Show that the correlation matrix of the output $\mathbf{Y}(t)$ is given by Equation 9.7-4.
(b) Derive the corresponding equation for matrix covariance functions.

9.45 In geophysical signal processing one often has to simulate a multichannel random process. The following problem brings out an important constraint on the power spectral density matrix of such a vector random process. Let the N-dimensional vector random process $\mathbf{X}(t)$ be WSS with correlation matrix

$$\mathbf{R}_{\mathbf{XX}}(\tau) \triangleq E[\mathbf{X}(t+\tau)\mathbf{X}^{\dagger}(t)]$$

and power spectral density matrix

$$\mathbf{S}_{\mathbf{XX}}(\omega) \triangleq FT\{\mathbf{R}_{\mathbf{XX}}(\tau)\}.$$

Here $FT\{\cdot\}$ denotes the matrix Fourier transform, that is, the (i,j)th component of $\mathbf{S}_{\mathbf{XX}}$ is the Fourier transform of the (i,j)th component of $\mathbf{R}_{\mathbf{XX}}$, which is $E[X_i(t+\tau)\, X_j^*(t)]$, where $X_i(t)$ is the ith component of $\mathbf{X}(t)$.

(a) For constants $a_1, \ldots, a_N$ define the WSS scalar process

$$Y(t) \triangleq \sum_{i=1}^{N} a_i X_i(t).$$

Find the power spectral density of $Y(t)$ in terms of the components of the matrix $\mathbf{S}_{\mathbf{XX}}(\omega)$.
(b) Show that the psd matrix $\mathbf{S}_{\mathbf{XX}}(\omega)$ must be a positive semidefinite matrix for each fixed ω; that is, we must have $\mathbf{a}^T\mathbf{S}_{\mathbf{XX}}(\omega)\mathbf{a}^* \geq 0$ for all complex column vectors $\mathbf{a}$.

9.46 Consider the linear system shown in Figure P9.46 excited by the two *orthogonal*, zero-mean, jointly WSS random processes $X(t)$, "the signal," and $U(t)$, "the noise." Then the input to the system G is

$$Y(t) = h(t) * X(t) + U(t),$$

which models a distorted-signal-in-noise estimation problem. If we pass this $Y(t)$, "the received signal" through the filter G, we get an estimate $\hat{X}(t)$. Finally $\varepsilon(t)$ can be thought of as the "estimation error"

$$\varepsilon(t) = \hat{X}(t) - X(t).$$

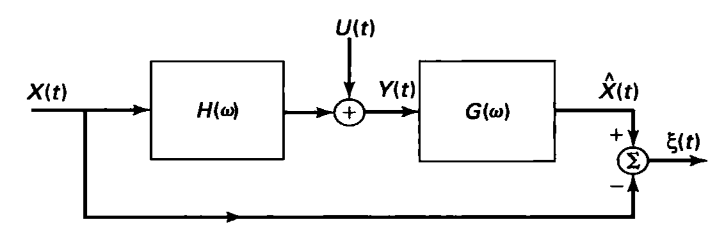


Figure P9.46 System for evaluating estimation error.

In this problem we will calculate some relevant power spectral densities and cross-power spectral density.

(a) Find $S_{YY}(\omega)$.
(b) Find $S_{X\hat{X}}(\omega) = S^*_{\hat{X}X}(\omega)$, in terms of H, G, S_{XX}, and S_{UU}.
(c) Find $S_{\varepsilon\varepsilon}(\omega)$.
(d) Use your answer to part (c) to show that to minimize $S_{\varepsilon\varepsilon}(\omega)$ at those frequencies where

$$S_{XX}(\omega) >> S_{UU}(\omega),$$

we should have $G \approx H^{-1}$ and where

$$S_{XX}(\omega) << S_{UU}(\omega)$$

we should have $G \approx 0$.

***9.47** Let $X(t)$, the input to the system in Figure P9.47, be a stationary Gaussian random process. The power spectral density of $Z(t)$ is measured experimentally and found to be

$$S_{ZZ}(\omega) = \pi\delta(\omega) + \frac{2\beta}{(\omega^2+\beta^2)(\omega^2+1)}.$$

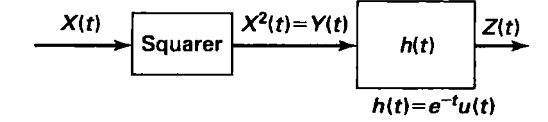


Figure P9.47 Squarer nonlinearity followed by linear filter.

(a) Find the correlation function of $Y(t)$ in terms of β.
(b) Find the correlation function of $X(t)$.

9.48 Consider the two-state Markov chain $N(t)$ shown in Figure P9.48, taking on values 1 and 2. While in state 1, the transition time to state 2 has average rate $\lambda_1 = 1$. In state 2, the transition time to state 1 has average rate $\lambda_2 = 2$. Denote the state probabilities as $P_1(t)$ and $P_2(t)$, where $P_i(t) \triangleq P[N(t) = i]$ for $i = 1, 2$.

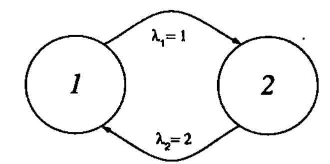


Figure P9.48 Two-state Markov chain state-transition diagram.

(a) Derive the differential equations for the $P_i(t)$.
(b) Find their steady-state solution.

9.49 The Schwarz inequality for complex-valued random variables states that

$$|E[XY^*]| \leq \sqrt{E[|X|^2]\, E[|Y|^2]}\ ,$$

for two random variables X and Y.

(a) Use the Schwarz inequality to derive the corresponding result for WSS random processes $X(t)$ and $Y(t)$,

$$|R_{XY}(\tau)| \leq \sqrt{R_{XX}(0)\, R_{YY}(0)}\ .$$

(b) Find the corresponding result for cross-power spectral densities,

$$|S_{XY}(\omega)| \leq \sqrt{S_{XX}(\omega)\, S_{YY}(\omega)}\ .$$

Hint: Interpret the result of part (a) in terms of cross- and auto-power spectra, and then introduce a narrow bandpass filter centered at an arbitrary frequency ω.

9.50 The Wiener process, also called Brownian motion, is the integral of white noise. Letting $B(t)$ denote the Wiener process, with $W(t)$ denoting the white noise, we can write

$$B(t) = \int_0^t W(\tau)d\tau, \qquad t \geq 0.$$

Take $W(t)$ to be a standard white noise with correlation function $R_W(\tau) = \delta(\tau)$.

(a) Find and sketch the cross-correlation function $R_{BW}(t_1, t_2)$.
(b) Find and sketch the autocorrelation function $R_{BB}(t_1, t_2)$.

9.51 Consider the two-processor reliability problem of Example 9.2-4 in the text, a three-state continuous-time Markov random process $X(t)$ with state-transition diagram shown in Figure P9.51. Here, $X(t)$ denotes the number of processors "up" at time t.

(a) Write the state probability vector $\mathbf{p}(t)$ differential equation

$$d\mathbf{p}(t)/dt = \mathbf{A}\mathbf{p}(t)$$

and explicitly find the generator matrix $\mathbf{A}$.

(b) We determine the steady-state probability vector $\mathbf{p}$ by solving the homogeneous matrix-vector equation $\mathbf{A}\mathbf{p} = \mathbf{0}$, subject to the constraint that all the probabilities in the probability vector $\mathbf{p}$ sum to 1. Someone claims that the "probability flows" across the dashed vertical lines in Figure P9.51 must balance in the steady-state, that is, $2\mu p_0 = \lambda p_1$ and $\mu p_1 = 2\lambda p_2$, where the p_i are the elements of the vector $\mathbf{p}$, that is, the steady-state probabilities of being in state i, $i = 0, 1, 2$. State why this is a reasonable assertion, and prove it by showing that the resulting equations satisfy $\mathbf{A}\mathbf{p} = \mathbf{0}$.

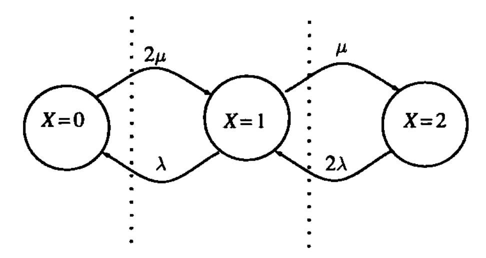


Figure P9.51

(c) Solve for the numerical steady-state probability values in the case when $\lambda = 0.001$ and $\mu = 0.1$ per hour.

9.52 Consider the three-input, two-output LSI system shown in Figure P9.52. The input random processes $X_1(t), X_2(t)$, and $U(t)$ are jointly WSS and pairwise orthogonal, that is, $X_1 \perp X_2, X_1 \perp U$, and $X_2 \perp U$. We are given the following functions: the indicated system functions H, G, and B, plus the three-input power spectral densities $S_{X_1X_1}, S_{X_2X_2}$, and S_{UU}. You may express your answers in terms of these functions.

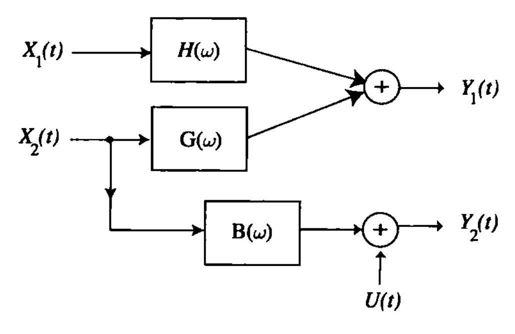


Figure P9.52

(a) Find the input/output cross-power spectral density $S_{Y_1X_1}(\omega)$.
(b) Find the input/output cross-power spectral density $S_{Y_2X_2}(\omega)$.
(c) Find the output cross-power spectral density $S_{Y_1Y_2}(\omega)$.

9.53 Consider the following tapped delay-line problem. We have a random sequence A_n for the taps and a WSS random process $X(t)$ as the signal model. Assume the total number of taps is N and the tap spacing is T. Assume also that the random sequence A_n and the random process $X(t)$ are jointly independent. The tapped delay-line output is therefore

$$Y(t) = \sum_{n=0}^{N-1} A_n X(t - nT).$$

The correlation function for the random sequence of tap weights is given as $R_A(n_1, n_2)$, and the correlation function of the WSS random process is given as $R_X(\tau)$.

(a) Find the output correlation function $R_Y(t_1, t_2)$ in terms of the given functions and parameters.
(b) Does the wide-sense stationarity of $Y(t)$ depend on whether the random sequence A_n is WSS? Justify your answer.
(c) In finding your result in part (a), is it sufficient that A_n and $X(t)$ be uncorrelated? Why?

9.54 Let a certain wireless packet channel (Gilbert channel model) having a *good* state and a *bad* state be modeled as a continuous-time, two-state Markov chain with transition rates as given in Figure P9.54.

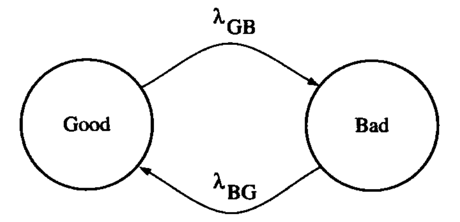


Figure P9.54 Gilbert channel model.

(a) Find the steady-state probability of being in the bad state.
(b) In the good state, all packets are received. In the bad state, all packets are lost. This leads to bursts or clusters of lost packets. In a *packet-loss burst*, all

packets are lost. What is the average length of a packet-loss burst? Justify. Note that the chain is in the bad state for the full duration of a packet-loss burst.

9.55 Consider a Poisson random process $N(t)$ with average arrival rate $\lambda = 3$.

(a) Find the probability that $N(4) = 2$.
(b) Find the joint probability that $N(1) = 1$ and $N(2) = 2$.

9.56 Consider the system shown in Figure P9.56.

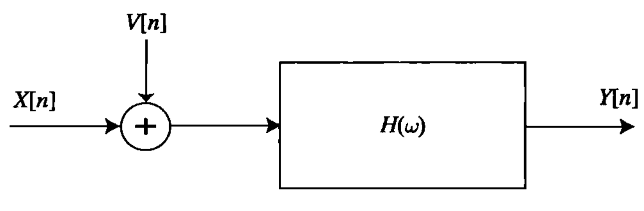


Figure P9.56

Let $X[n]$ and $V[n]$ be WSS and mutually uncorrelated with zero means and power spectral densities $S_{XX}(\omega)$ and $S_{VV}(\omega)$, respectively.

(a) Find the psd of the output $Y[n]$.
(b) Find the cross-power spectral density between input $X(t)$ and output $Y(t)$, that is, $S_{XY}(\omega)$.

9.57 Let $X(t)$ and $Y(t)$ be two zero-mean random processes with known correlation coefficient function

$$\rho_{XY}(t_1, t_2) \triangleq \frac{E[X(t_1)Y^*(t_2)]}{\sqrt{E[|X(t_1)|^2]E[|Y(t_2)|^2]}},$$

and assume that the average powers $E[|X(t)|^2] = E[|Y(t)|^2] \triangleq P$, a constant. Next, add two random noises $U(t)$ and $V(t)$, jointly orthogonal to $X(t)$ and $Y(t)$,

$$\widetilde{X}(t) \triangleq X(t) + U(t),$$

$$\widetilde{Y}(t) \triangleq Y(t) + V(t),$$

where U and V are also orthogonal to each other and of zero mean, and with average powers $E[|U(t)|^2] = E[|V(t)|^2] \triangleq \epsilon$, a constant. Find the correlation coefficient function of the tilde processes, that is, $\rho_{\widetilde{X}\widetilde{Y}}(t_1, t_2)$ in terms of that of the original processes X and Y.

9.58 Consider the system shown in Figure P9.58. The two-input random sequences are WSS and given in terms of their power spectral densities:

$$S_{XX}(\omega) = \frac{1}{\omega^2 + 5} \quad \text{and} \quad S_{VV}(\omega) = \frac{2\omega^2 + 8}{(\omega^2 + 3)(\omega^2 + 5)}.$$

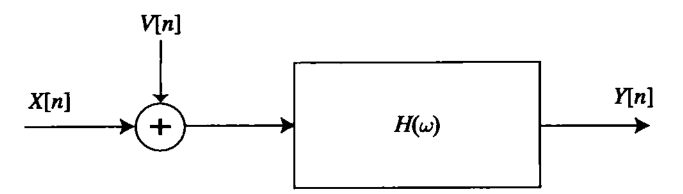


Figure P9.58 System with signal plus noise input.

The system function $H(\omega)$ is given as $10[u(\omega+\pi/2)-u(\omega-\pi/2)]$ over the interval $[-\pi,+\pi]$, where the function u is the unit step. Assume that X and V are zero mean.

(a) Assuming X and V are uncorrelated, find the psd of the output random sequence $Y[n]$.

(b) Let the cross-power spectral density of X and V be specified as

$$S_{XV}(\omega)=\frac{1}{\omega^2+5},$$

and find the new output power spectral density of Y.

9.59 Consider the random process $X(t)=\cos(\omega_0 t+\Theta)$, where Θ is a random variable uniformly distributed over the interval $[0,2\pi]$, and ω_0 is a fixed frequency. Find the first-order pdf $f_X(x;t)$. Is the process stationary of first order? Find the conditional pdf of $X(t_2)$ given $X(t_1)=x_1$, which we denote by $f_X(x_2|x_1;t_1,t_2)$. You may assume $t_1<t_2$.

***9.60** Let $Z(t)=X(t)+jY(t)$, where $X(t)$ and $Y(t)$ are jointly WSS and *real-valued* random processes. Assume that $X(t)$ and $Y(t)$ are mutually orthogonal with zero-mean functions. Define a new random process in terms of a modulation to a carrier frequency ω_0 as $U(t)=Re\{Z(t)e^{-j\omega_0 t}\}$. Given the relevant correlation functions, that is, $R_{XX}(\tau)$ and $R_{YY}(\tau)$, find general conditions on them such that $U(t)$ is also a WSS random process. Show that your conditions work, that is, that the resulting process $U(t)$ is actually WSS. Some helpful trigonometric identities:

$$\cos(\alpha\pm\beta)=\cos\alpha\cos\beta\mp\sin\alpha\sin\beta$$

$$\sin(\alpha\pm\beta)=\sin\alpha\cos\beta\pm\cos\alpha\sin\beta.$$

9.61 Find the steady-state probabilities of the four-state Markov chain shown in Figure P9.61. Express your answers in terms of the exponential rates λ_i and μ_i. Note the state labels are conveniently given as 1 through 4.

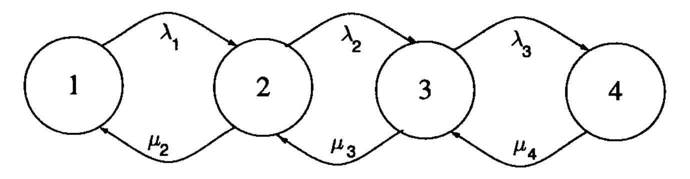


Figure P9.61

Hint: Remember the probability flow concept from Problem 9.51 (b).

9.62 Consider the three-state Markov process $X(t)$ with state-transition diagram shown in Figure P9.62. Here the state labels are the actual outputs, that is, $X(t) = 3$ all the while the process is in state 3. The state transitions are governed by jointly independent, exponentially distributed interarrival times, with average rates as indicated on the branches

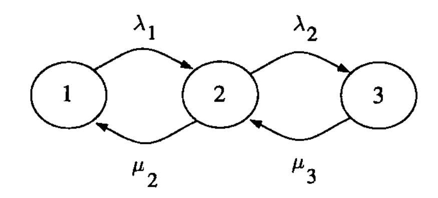


Figure P9.62 Three-state Markov process.

(a) Given that we start at state 2 at time $t = 0$, what is the probability that we leave this state for the first time at time t, for some arbitrary $t > 0$?

(b) Find the vector differential equation for the state probability at time $t \geq 0$,

$$\frac{d\mathbf{p}}{dt} = \mathbf{A}\mathbf{p}(t),$$

where $\mathbf{p}(t) = [p_1, p_2, p_3]^T$, expressing the generator matrix $\mathbf{A}$ in terms of the λ_i and μ_i.

(c) Show that the solution for $t \geq 0$ can be expressed as

$$\mathbf{p}(t) = \exp(\mathbf{A}t)\ \mathbf{p}(0),$$

where $\mathbf{p}(0)$ is the initial probability vector and the matrix $\exp(\mathbf{A}t)$ is defined by the infinite series

$$\exp(\mathbf{A}t) \triangleq \mathbf{I} + \mathbf{A}t + \frac{1}{2!}(\mathbf{A}t)^2 + \frac{1}{3!}(\mathbf{A}t)^3 + \frac{1}{4!}(\mathbf{A}t)^4 + \cdots$$

Do not worry about convergence of this series, but it is known that it absolutely converges for all finite t.

REFERENCES

9-1. S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes.* New York: Academic Press, 1975.

9-2. L. Kleinrock, *Queueing Systems*, Vol. 1: *Theory.* New York: John Wiley, 1975.

9-3. T. Kailath, *Linear Systems.* Upper Saddle River, NJ: Prentice Hall, 1980.

9-4. E. W. Kamen, *Introduction to Signals and Systems*, 2nd edition. New York: Macmillan, 1990, p. 172.

9-5. E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems.* New York: Springer-Verlag, 1985, pp. 62–63.
9-6. P. Billingsley, *Probability and Measure*, New York: John Wiley, 1979, pp. 467–477.
9-7. J. L. Doob, *Stochastic Processes.* New York: John Wiley, 1953.
9-8. P. Whittle, "On Stationary Process in the Plane," *Biometrika*, Vol. 41 (1954), pp. 434–449.
9-9. A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd Edition. Upper Saddle River, NJ: Prentice Hall, 1999, Chapters 2–3.

APPENDIX A Review of Relevant Mathematics

This section will review the mathematics needed for the study of probability and random processes. We start with a review of basic discrete and continuous mathematical concepts.

A.1 BASIC MATHEMATICS

We review the concept of sequence and present several examples. We then look at summation of sequences. Next the Z-transform is reviewed.

Sequences

A sequence is simply a mapping of a set of integers into the set of real or complex numbers. Most often the set of integers is the nonnegative integers $\{n \geq 0\}$ or the set of all integers $\{-\infty < n < +\infty\}$.

An example of a sequence often encountered is the exponential sequence a^n for $\{n \geq 0\}$, which is plotted in Figure A.1-1 for several values of the real number a. Note that for $|a| > 1$, the sequence diverges, while for $|a| < 1$, the sequence converges to 0. For $a = 1$, the sequence is the constant 1, and for $a = -1$, the sequence alternates between $+1$ and -1.

A related and important sequence is the complex exponential $\exp(j\omega n)$. These sequences are eigenfunctions of linear time-invariant systems, which just means that for such a system with frequency response function $H(\omega)$, the response to the input $\exp(j\omega n)$ is just $H(\omega)\exp(j\omega n)$, a scaled version of the input.

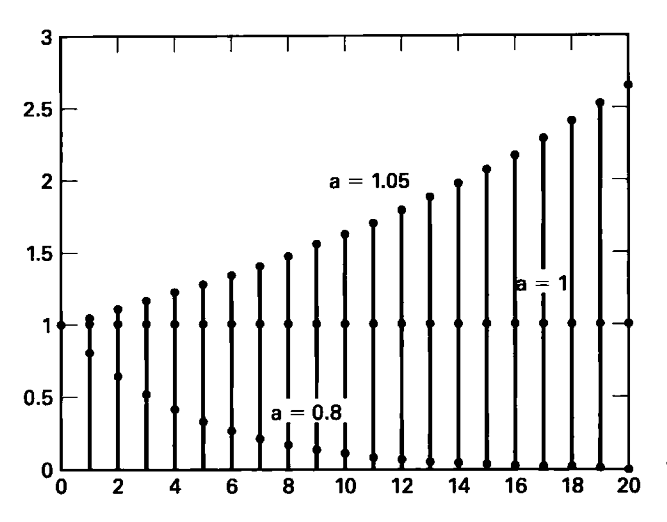


Figure A.1-1 Plot of exponential sequence for three values of $a = 1.05, 1.0$, and 0.8.

Convergence

A sequence, denoted $x[n]$ or x_n, which is defined on the positive integers $n \geq 1$, converges to a limiting value x if the values $x[n]$ become nearer and nearer to x as n becomes large. More precisely, we can say that for any given $\varepsilon > 0$, there must exist a value $N_0(\varepsilon)$ such that for all $n \geq N_0$, we have $|x[n] - x| < \varepsilon$. Note that N_0 is allowed to depend on ε.

Example A.1-1

Let the sequence a_n be given as

$$a_n = 2^n/(2^n + 3^n),$$

and find the limit as $n \to \infty$. From observation, we see that the limit is $a_n = 0$. To complete the argument, we can then express $N_0(\varepsilon)$ from the equation

$$\left|\frac{2^n}{2^n + 3^n}\right| < \varepsilon$$

as

$$N_0(\varepsilon) = \frac{\ln\left(\frac{1-\varepsilon}{\varepsilon}\right)}{\ln\frac{3}{2}},$$

where we assume that $0 < \varepsilon < 1$. We note that for any fixed $0 < \varepsilon < 1$, the value N_0 is finite as required.

Summations

Summations of sequences arise quite often in our work. A common sequence used to illustrate summation concepts is the geometric sequence a^n. The following summation formula can be readily derived: Take $n_2 \geq n_1$.

$$\sum_{n=n_1}^{n_2} a^n = \frac{a^{n_1} - a^{n_2+1}}{1-a} \quad \text{for} \quad a \neq 1. \tag{A.1-1}$$

Of course, when $a = 1$, the summation is just $n_2 - n_1 + 1$. A simple way to see the validity of Equation A.1-1 is to first define $S = \sum_{n_1}^{n_2} a^n$ and then note that, by the special property of the geometric sequence,

$$aS = S + a^{n_2+1} - a^{n_1}.$$

Then, by solving for S, we derive Equation A.1-1 when $a \neq 1$.

When $|a| \leq 1$, the upper limit of summation can be extended to ∞ to yield

$$\sum_{n=n_1}^{\infty} a^n = \frac{a^{n_1}}{1-a} \quad \text{for} \quad |a| < 1. \tag{A.1-2}$$

Another useful related summation is:

$$\sum_{n=n_1}^{\infty} na^n = \frac{n_1 a^{n_1}(1-a) + a^{n_1+1}}{(1-a)^2} \quad \text{for} \quad |a| < 1. \tag{A.1-3}$$

Equations A.1-2 and A.1-3 most often occur with $n_1 = 0$.

Z-Transform

This transform is very helpful in solving for various quantities in a linear time-invariant system and also for the solution of linear constant-coefficient difference equations. The Z-transform is defined for a deterministic sequence $x[n]$ as follows:

$$\mathsf{X}(z) = \sum_{n=-\infty}^{+\infty} x[n]z^{-n}, \text{ for } z \in \mathscr{R}.$$

In this equation, the region $\mathscr{R}$ is called the *region of convergence* and denotes the set of complex numbers z for which the transform is defined. This set $\mathscr{R}$ is further specified as those z for which the relevant sum converges absolutely, that is,

$$\sum_{n=-\infty}^{+\infty} |x[n]||z|^{-n} < \infty.$$

This region $\mathscr{R}$ can be written in general as $\mathscr{R} = \{z : R_- < |z| < R_+\}$, an annular shaped region. The set $\{z | R_- < |z| < R_+\}$ is to be read as "the set of all points z whose magnitude (length) is greater than R_- and less than R_+."

Example A.1-2

Let the discrete-time sequence $x[n]$ be given as the exponential

$$x[n] = a^n \exp(j\omega_0 n)u[n],$$

where $u[n]$ denotes the unit step sequence, $u[n] = 1$ for $n \geq 0$ and $u[n] = 0$ for $n < 0$.

Calculating the Z-transform, we get

$$\begin{aligned} \mathsf{X}(z) &= \sum_{n=0}^{\infty} a^n \exp(j\omega_0 n) z^{-n} \\ &= \sum_{n=0}^{\infty} (ae^{j\omega_0} z^{-1})^n \qquad \text{(A.1-4)} \\ &= \frac{1}{1 - ae^{j\omega_0} z^{-1}} \quad \text{for} \quad |z| > |a|. \quad \mathscr{R} = \{z : |a| < |z|\}. \end{aligned}$$

The Z-transform is quite useful in discrete-time signal processing because of the following fundamental theorem relating convolution and multiplication of the corresponding Z-transforms.

Theorem A.1-1 Consider the convolution of two absolutely summable sequences $x[n]$ and $h[n]$, which generates a new sequence $y[n]$ as follows:

$$y[n] = \sum_{m=-\infty}^{+\infty} x[m]h[n-m]$$

which we denote operationally as $y = h * x$. Then the Z-transform of $y[n]$ is given in terms of the corresponding Z-transforms of x and h as

$$\mathsf{Y}(z) = \mathsf{H}(z)\mathsf{X}(z) \quad \text{for} \quad z \in \mathscr{R}_h \cap \mathscr{R}_x.$$

Because the two sequences h and x are absolutely summable, their regions of convergence $\mathscr{R}_h$ and $\mathscr{R}_x$ will both include the unit circle of the z-plane, that is, $\{|z| = 1\}$. Then the Z-transform $\mathsf{Y}(z)$ will exist for $z \in \mathscr{R}_h \cap \mathscr{R}_x$, which is then guaranteed to be nonempty. ■

After obtaining the Z-transform of a convolution using this result, one can often take the inverse Z-transform to get back the output sequence $y[n]$. There are several ways to do this, including expansion of the Z-transform $\mathsf{Y}(z)$ in a power series, doing long division in the typical case when $\mathsf{Y}(z)$ is a ratio of polynomials in z, and the most powerful method, the method of residues. This last method, along with the residue method for inverse Laplace transforms, is the topic of Section A.3 of this appendix.

A.2 CONTINUOUS MATHEMATICS

The intent here is to review some ideas from the integral calculus of one- and two-dimensional functions of real variables.

Definite and Indefinite Integrals

In a basic calculus course, we study two types of integrals, definite and indefinite:

$$\int x^2 dx = \frac{1}{3}x^3 + C \qquad \text{indefinite,}$$

$$\int_a^b x^2 dx = \frac{1}{3}b^3 - \frac{1}{3}a^3 \qquad \text{definite.}$$

In this course we will most always write the definite integral, almost never the indefinite integral. This is because we will use integrals to measure specific quantities, not merely to determine the class of functions that have a given derivative. Please note the difference between these two integrals. Unlike the indefinite integral, the definite integral is a function of its upper and lower limits, but not of x itself! Sometimes we refer to x in our definite integrals as a "dummy variable" for this reason, that is, x could just as well be replaced by another variable, say y, with no change resulting to our definite integral, that is,

$$\int_a^b x^3 dx = \int_a^b y^3 dy.$$

To compute the definite integral we first compute the indefinite integral and then subtract its evaluation at the lower limit from its evaluation at the upper limit.

In elementary calculus courses it is often not stressed that definite integrals are operations on sets and that there are integrals that are not associated with the "area under a curve," that is, so-called Riemann integrals. Consider the definite integral

$$I = \int_a^b f(x)dz(x).$$

Here the set of points is $\{x : a \leq x \leq b\}$ and the integral is computed by assigning numerical values to the points in an n-partition of the interval (a, b) vis-à-vis $\Delta x \equiv (b-a)/n$ in the set, for example

$$I_n(a,b) = \sum_i f(i\Delta x) \times \Big(z(i\Delta x + \Delta x/2) - z(i\Delta x - \Delta x/2)\Big) \xrightarrow[\substack{\Delta x \to 0,\\ n \to \infty}]{} I$$

where $i\Delta x, i\Delta x \pm \Delta x/2 \in \{x : a \leq x \leq b\}$. If $z(x) = x$, then I becomes the well-known "area under the curve" Riemann integral. But in some cases the Riemann integral won't suffice. For example, consider the expectation operation we encountered in Chapter 4, that is, $E[X] = \int_{-\infty}^{\infty} x f_x(x)dx$, which will converge to the desired result if $f_X(x)$ is well-defined, that is, a bounded function with only a finite set of discontinuities etc. But if $f_X(x)$ does not fall into this category of functions, we can still compute $E[X]$ from the integral $E[X] = \int_{-\infty}^{\infty} x dF_X(x)$, where $F_X(x)$ is the CDF of X. This type of integral is called a Stieljes integral and is a generalization of the "area under the curve" type integral that is taught in beginning calculus courses. For example, if $F_X(x) = (1 - e^{-\lambda x})u(x)$, then $dF_X(x) = (\lambda e^{-\lambda x})u(x)dx$ and $E[X] = \int_0^{\infty} x\lambda e^{-\lambda x}dx = 1/\lambda$.

Differentiation of Integrals

From time to time, it becomes necessary to differentiate an integral with respect to a parameter which appears either in the upper limit, the lower limit, or the integrand itself:

$$\frac{d}{dy}\int_{a(y)}^{b(y)} f(x,y)dx = f(b(y),y)\frac{db(y)}{dy} - f(a(y),y)\frac{da(y)}{dy} + \int_{a(y)}^{b(y)} \frac{\partial f(x,y)}{\partial y}dx. \qquad \text{(A.2-1)}$$

This important formula is derived by recalling that for a function $I(b,a,y)$ where in turn, $b = b(y)$ and $a = a(y)$ are two functions of y, we have

$$\frac{dI}{dy} = \frac{\partial I}{\partial b}\frac{db}{dy} + \frac{\partial I}{\partial a}\frac{da}{dy} + \frac{\partial I}{\partial y}.$$

If we denote

$$I \triangleq \int_{a(y)}^{b(y)} f(x,y)dx$$

and define a function $F(x,y)$ such that

$$f(x,y) \triangleq \frac{\partial F(x,y)}{\partial x},$$

then clearly

$$\frac{\partial I}{\partial b} = f(b(y),y)$$

$$\frac{\partial I}{\partial a} = -f(a(y),y)$$

$$\frac{\partial I}{\partial y} = \frac{\partial}{\partial y}\int_{a(y)}^{b(y)} f(x,y)dx = \int_{a(y)}^{b(y)} \frac{\partial}{\partial y} f(x,y)dx.$$

The last step on the right follows from treating $b(y)$ and $a(y)$ as *constants*, since the variation of I arising from its upper and lower limits is already counted by the first two terms.

An example of use of this formula, which arises in the study of how systems transform probability functions, is shown next.

Example A.2-1

Consider the example where the function $f(x,y) = x + 2y$,

$$\frac{\partial}{\partial y}\int_0^y (x+2y)^2 dx = (y+2y)^2 1 - (0+2y)^2 0 + \int_0^y 4(x+2y)dx$$

$$= (3y)^2 + 4\left(\frac{1}{2}x^2 + 2yx\right)\Big|_0^y$$

$$= (3y)^2 + 2y^2 + 8y^2 = 19y^2$$

Integration by Parts

Integration by parts is a useful technique for explicit calculation of integrals. We write the formula as follows:

$$\int_a^b u(x)dv(x) = u(x)v(x)|_a^b - \int_a^b v(x)du(x), \tag{A.2-2}$$

where u and v denote functions of the variable x with the integral extending over the range $a \leq x \leq b$. This formula is derived using the chain rule for derivatives, applied to the derivative of the product function $u(x)v(x)$. An example is shown below. Integration by parts is useful to extend the class of integrals that are doable analytically.

Example A.2-2

Consider the following integration problem:

$$\int_0^\infty xe^{-2x}dx$$

Let $u(x) = x$ and $dv(x) = e^{-2x}dx$; then using the above integration by parts formula we obtain

$$\begin{aligned}\int_0^\infty xe^{-2x}dx &= x\left(-\frac{1}{2}e^{-2x}\right)\Big|_0^\infty - \int_0^\infty \left(-\frac{1}{2}e^{-2x}\right)dx \\ &= \frac{1}{2}\int_0^\infty e^{-2x}dx = \frac{1}{2}\left(-\frac{1}{2}e^{-2x}\right)\Big|_0^\infty \\ &= \frac{1}{4}.\end{aligned}$$

Completing the Square

The method of completing the square is applied to the calculation of integrals by transforming an unknown integral into a known one by turning the argument of its integrand into a perfect square. For example, consider making a perfect square out of $x^2 + 4x$. We can transform it into the perfect square $(x+2)^2$ by adding and subtracting 4, that is,

$$x^2 + 4x = (x+2)^2 - 4.$$

To see how this polynomial concept can be used to calculate integrals, consider the well-known Gaussian integral that we often encounter in this course:

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2}dx = \sqrt{2\pi}.$$

If, instead we need to calculate

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2+4x)}dx = ?,$$

we can do so by completing the square as follows:

$$e^2 \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2+4x+4)} dx$$

where we have multiplied by e^{-2} inside the integral and by e^2 outside. Then we continue,

$$= e^2 \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x+2)^2} dx.$$

With the change of variables $y = x + 2$, this then becomes

$$= e^2 \int_{-\infty}^{+\infty} e^{-\frac{1}{2}y^2} dy$$

$$= e^2\sqrt{2\pi}.$$

Double Integration

Integrals on the (x, y) plane are properly called double integrals. The infinitesimal element is an area, written as $dxdy$. We often evaluate these integrals in some order, say x first and then y, or vice versa. Then the integral is called an iterated integral. We can write the three possible situations as follows:

$$\int_{y_1}^{y_2} \left(\int_{x_1}^{x_2} f(x, y)dx \right) dy = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)dx\, dy = \int_{x_1}^{x_2} \left(\int_{y_1}^{y_2} f(x, y)dy \right) dx,$$

where the integral in the middle is the true double or area integral. Since limiting operations are the basis for any integral, there is actually a question of whether the three two-dimensional integrals are always equal. Fortunately, an advanced result in measure theory [9-1] shows that when the integrals are defined in the modern Lebesgue sense, then all three either exist and are equal, or do not exist. We will consider only the ordinarily occurring case where the above three integrals exist and are equal.

Note that on the left, when we integrate on x first, that the limits are interchanged versus the situation on the right where we integrate in the y direction first. The double or area integral in the middle, adopts the notation that one reads the limits in x, y order, just as in the function arguments and the area differential $dxdy$. Thus, there should be no confusion in interpreting such expressions as

$$\int_1^3 \int_0^5 xe^{-y} dx\, dy,$$

since we would read this correctly as an integral over the rectangle with opposite corners $(x, y) = (1, 0)$ and $(x, y) = (3, 5)$.

Functions

A function is a unique mapping from a domain space $\mathscr{X}$ to a range space $\mathscr{Y}$. The only condition is uniqueness which means that only one y goes with each x, that is, $f(x)$ has one and only one value. An example is $f(x) = x^2$. A counterexample is $f(x) = \pm\sqrt{x}$.

Monotone Functions. A monotone function of the real variable x is one that *always* increases as x increases or *always* decreases as x increases. The former, with the positive slope, is called *monotone increasing*, while the latter, with the negative slope, is called *monotone decreasing*, as illustrated in Figures A.2-1 and A.2-2. If a function is monotone except for some flat regions of zero slope, then we use the terms monotone nondecreasing or monotone nonincreasing to describe them, as illustrated in Figure A.2-3.

Inverse Functions. A function may or may not have an inverse. The inverse function exists when the original function has the additional uniqueness property that to each y in $\mathscr{Y}$, there corresponds only one x (in $\mathscr{X}$). This allows us to define an inverse function $f^{-1}(y)$ to map



Figure A.2-1 Example of a *monotone increasing* function.

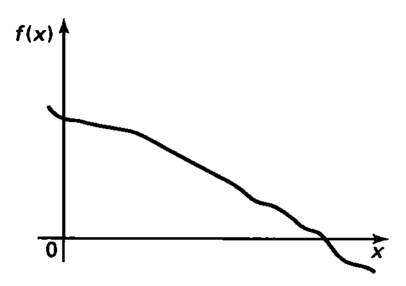


Figure A.2-2 Example of a *monotone decreasing* function.

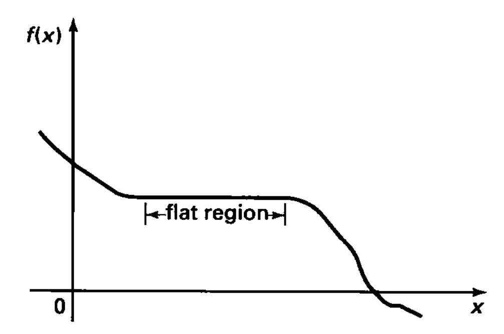


Figure A.2-3 Example of a monotone *nonincreasing* function.

back from $\mathscr{Y}$ to $\mathscr{X}$. We note that a sufficient condition for the inverse function to exist is that the original function $f(x)$ is monotone increasing *or* monotone decreasing. The function sketched in Figure A.2-3 does not have an inverse due to the flat section of zero slope.

A.3 RESIDUE METHOD FOR INVERSE FOURIER TRANSFORMATION†

In Chapters 8 and 9, we defined the power spectral density (psd) $S(\omega)$ for both discrete and continuous time and showed that the psd is central to analyzing LSI systems with random sequence and process inputs. We often want to take an inverse transform to find the correlation function corresponding to a given psd to obtain a time-domain characterization. This section summarizes the powerful residue method for accomplishing the necessary inverse Fourier transformation.

We start by recalling the relation between the psd and correlation function for a WSS random process,

$$S(\omega) = \int_{-\infty}^{+\infty} R(\tau)e^{-j\omega\tau}d\tau,$$

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S(\omega)e^{+j\omega\tau}\, d\tau.$$

To apply the residue method of complex variable theory [A-3] to the evaluation of the above IFT, we must first express this integral as an integral along a contour in the *complex s-plane.* We define a new function S of the *complex variable* $s = \sigma + j\omega$ as follows.

First we define $\mathsf{S}(s)$ on the imaginary axis in terms of the function of a *real variable* $S(\omega)$ as

$$\mathsf{S}(s)|_{s=j\omega} \triangleq S(\omega).$$

Then we replace $j\omega$ by s to extend the function $\mathsf{S}(j\omega)$ to the entire complex plane. Thus,

$$\mathsf{S}(s)|_{s=j\omega} = S(\omega) = \int_{-\infty}^{+\infty} R(\tau)e^{-j\omega\tau}\, d\tau$$

so

$$\mathsf{S}(s) = \int_{-\infty}^{+\infty} R(\tau)e^{-s\tau}\, d\tau, \tag{A.3-1}$$

which is the two-sided Laplace transform of the correlation function R. Also by inverse Fourier transform,

$$\begin{aligned} R(\tau) &= \frac{1}{2\pi j}\int_{-\infty}^{+\infty} \mathsf{S}(s)|_{s=j\omega}e^{j\omega\tau}d(j\omega), \\ &= \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty} \mathsf{S}(s)e^{s\tau}\, ds, \end{aligned} \tag{A.3-2}$$

which is an integral along the imaginary axis of the s-plane.

†This material assumes that the reader is familiar with the discussions in Chapters 8 and 9.

The integral in Equation A.3-2 is called a *contour integral* in the theory of functions of a complex variable [A-2] [A-3], where it is shown that one can evaluate such an integral over a closed contour by the *method of residues.* This method is particularly easy to apply when the functions are rational; that is, the function is the ratio of two polynomials in s. Since this situation often occurs in linear systems whose behavior is modeled by differential equations, this method of evaluation can be very useful. We state the main result as a fact from the theory of complex variables.

Fact

Let $\mathsf{F}(s)$ be a function of the complex variable s, which is analytic inside and on a closed *counterclockwise* contour C except at P poles located inside C. The contour C encircles the origin. The P poles are located at $s = p_i, i = 1, \ldots, P$. Then

$$\frac{1}{2\pi j}\oint_C \mathsf{F}(s)ds = \sum_{\substack{p_i \text{ inside} \\ C}} \text{Res}[\mathsf{F}(s); s = p_i], \tag{A.3-3}$$

where

1. at a first-order pole, $\text{Res}[\mathsf{F}(s); s = p] = [\mathsf{F}(s)(s-p)]|_{s=p}$;
2. at a second-order pole, $\text{Res}[\mathsf{F}(s); s = p] = \frac{d}{ds}[\mathsf{F}(s)(s-p)^2]|_{s=p}$; and at an nth order pole
3. $\text{Res}[\mathsf{F}(s); s = p] = \frac{1}{(n-1)!}\left(\frac{d^{(n-1)}}{ds^{(n-1)}}[\mathsf{F}(s)(s-p)^n]\right)\Big|_{s=p}$.

In applying these results to our problem we first have to close the contour in some fashion. If we close the contour with a half-circle of infinite radius C_L as shown in Figure A.3-1, then

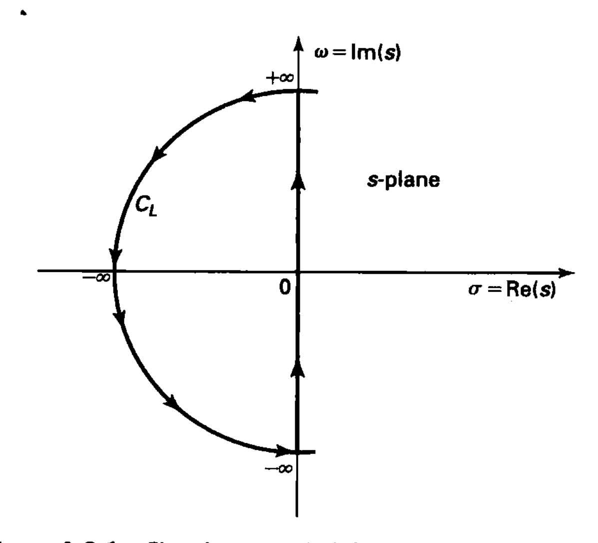


Figure A.3-1 Closed contour in left-half of s-plane for $\tau > 0$.

provided that the function being integrated, $\mathsf{S}(s)e^{s\tau}$, tends to zero fast enough as $|s| \to +\infty$, the value of the integral will not be changed by this closing of the contour. In other words, the integral over the semicircular part of the contour will be zero. The conditions for this are $|\mathsf{S}(s)|$ stays bounded as $|s| \to +\infty$, and

$$|e^{s\tau}| \to 0 \quad \text{as} \quad Re(s) \to -\infty,$$

the latter of which is satisfied for all $\tau > 0$. Thus, for positive τ we have

$$R(\tau) = \frac{1}{2\pi j} \oint_{C_L} \mathsf{S}(s)e^{s\tau} ds = \sum_{\substack{p_i \\ \text{inside } C_L}} \text{Res}[\mathsf{S}(s)e^{s\tau}; s = p_i],$$

Similarly, for $\tau < 0$ one can show that it is permissible to close the contour to the right as shown in Figure A.3-2, in which case we have

$$|e^{s\tau}| \to 0 \quad \text{as} \quad Re(s) \to +\infty,$$

so that we get

$$R(\tau) = \frac{1}{2\pi j} \oint_{C_R} \mathsf{S}(s)e^{s\tau} ds = - \sum_{\substack{p_i \\ \text{inside } C_R}} \text{Res}[\mathsf{S}(s)e^{s\tau}; s = p_i]$$

for $\tau < 0$, the minus sign arising from the clockwise traversal of the contour.

Example A.3-1

(*first-order psd*) Let

$$S(\omega) = 2\alpha/(\alpha^2 + \omega^2), \qquad 0 < \alpha < 1.$$

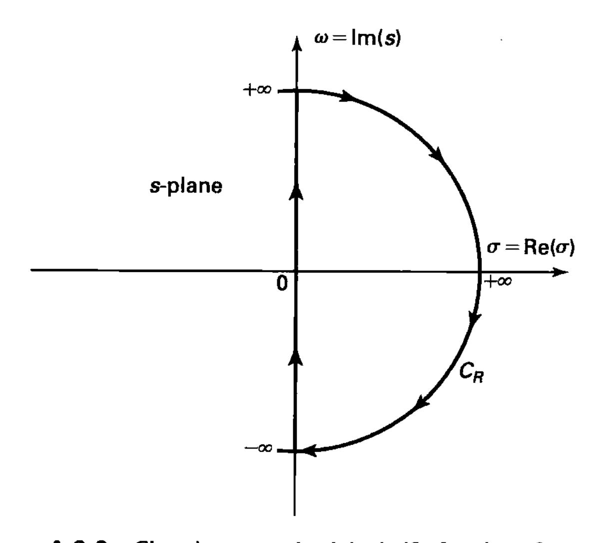


Figure A.3-2 Closed contour in right-half of s-plane for $\tau < 0$.

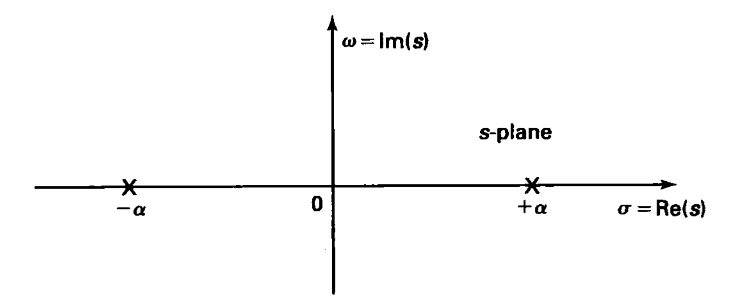


Figure A.3-3 Pole-zero diagram.

Then

$$\mathsf{S}(s)|_{s=j\omega} = S(\omega) = 2\alpha/(\alpha^2+\omega^2) = 2\alpha/(j\omega+\alpha)(-j\omega+\alpha),$$

so

$$\mathsf{S}(s) = \frac{2\alpha}{(s+\alpha)(-s+\alpha)},$$

where the configuration of the poles in the s-plane is shown in Figure A.3-3.

Evaluating the residues for $\tau > 0$, we get

$$R(\tau) = \text{Res}[\mathsf{S}(s)e^{s\tau}; s=-\alpha] = \left.\frac{2\alpha e^{s\tau}}{(-s+\alpha)}\right|_{s=-\alpha}$$
$$= \frac{2\alpha}{2\alpha}e^{-\alpha\tau},$$

while for $\tau < 0$ we get

$$R(\tau) = -\text{Res}[\mathsf{S}(s)e^{s\tau}; s=+\alpha]$$
$$= -\left.\frac{2\alpha e^{s\tau}(s-\alpha)}{(s+\alpha)(-s+\alpha)}\right|_{s=+\alpha}$$
$$= \left.\frac{-2\alpha e^{s\tau}}{(s+\alpha)(-1)}\right|_{s=\alpha} = \frac{2\alpha}{2\alpha}e^{\alpha\tau}.$$

Combining the results into a single formula, we get

$$R(\tau) = \exp(-\alpha|\tau|), \qquad -\infty < \tau < +\infty.$$

Inverse Fourier Transform for psd of Random Sequence

In the case of a random sequence one can do a similar contour integral evaluation in the complex z-plane. We recall the transform and inverse transform for a sequence:

$$S(\omega) = \sum_{m=-\infty}^{+\infty} R[m]e^{-j\omega m},$$

$$R[m] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} S(\omega)e^{+j\omega m}d\omega.$$

We rewrite the latter integral as a contour integral around the unit circle in a complex plane by defining the function of a complex variable, $\mathsf{S}(z)|_{z=e^{j\omega}} \triangleq S(\omega)$, and then substituting z for $e^{j\omega}$ into this new function to obtain the psd as a z-transform,

$$\mathsf{S}(z) = \sum_{m=-\infty}^{+\infty} R[m]z^{-m} \quad \text{and}$$

$$R[m] = \frac{1}{2\pi j}\oint_C \mathsf{S}(z)z^{m-1}dz \quad \text{where } C = \{|z| = 1\}. \tag{A.3-4}$$

In this case the contour is already closed and it encircles the origin in a counterclockwise direction, so we can apply Equation A.3-3 directly to obtain

$$R[m] = \sum_{\substack{p_i \\ \text{inside } C}} \text{Res}[\mathsf{S}(z)z^{m-1}; z = p_i],$$

where the sum is over the residues at the poles inside the unit circle. This formula is valid for all values of the integer m; however, it is awkward to evaluate for negative m due to the variable-order pole contributed by z^{m-1} at $z = 0$. Fortunately, a transformation mapping z to $1/z$ conveniently solves this problem, and we have [A-1],

$$\begin{aligned} R[m] &= \frac{1}{2\pi j}\oint_C \mathsf{S}(z^{-1})z^{-m+1}(-z^{-2}dz), \\ &= \frac{1}{2\pi j}\oint_C \mathsf{S}(z^{-1})z^{-m-1}dz, \end{aligned}$$

avoiding the variable-order pole for $m < 0$. We thus arrive at the prescription:
For $m \geq 0$

$$R[m] = \sum_{\substack{i\text{:poles} \\ \text{inside unit circle}}} \text{Res}[\mathsf{S}(z)z^{m-1}; z = p_i],$$

and for $m < 0$

$$R[m] = \sum_{\substack{i\text{:poles} \\ \text{outside unit circle}}} \text{Res}[\mathsf{S}(z^{-1})z^{-m-1}; z = p_i^{-1}].$$

Example A.3-2

(*first-order psd of random sequence*) We consider a psd given as

$$S(\omega) = \frac{2(1-\rho^2)}{(1+\rho^2) - 2\rho\cos\omega}, \qquad |\omega| \leq \pi, \tag{A.3-5}$$

which is plotted in Figure A.3-4.

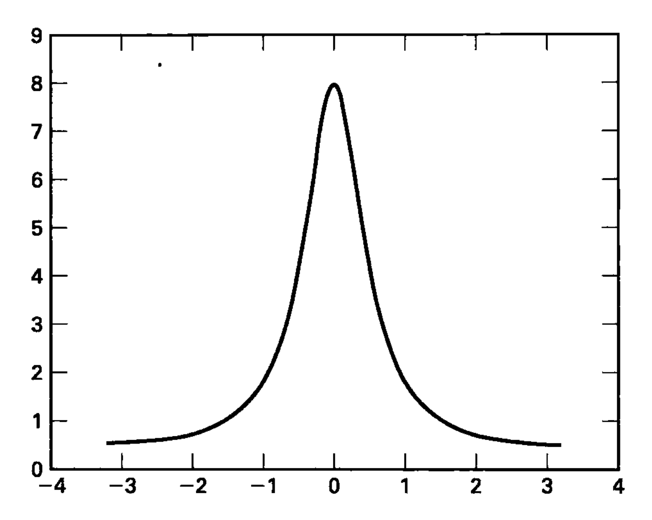


Figure A.3-4 Plot of psd $S(\omega)$ for a ρ value in (0,1).

Using the identify $\cos\omega = 0.5(\exp j\omega + \exp -j\omega)$, we can make this substitution in Equation A.3-5 to obtain the function of a complex variable,

$$\mathsf{S}(z)|_{z=e^{j\omega}} = S(\omega) = \frac{2(1-\rho^2)}{(1+\rho^2) - 2\rho\cos\omega}$$

$$= \frac{2(1-\rho^2)}{(1+\rho^2) - \rho(e^{+j\omega} + e^{-j\omega})}.$$

Then we replace $e^{j\omega}$ by z to obtain the function of z,

$$\mathsf{S}(z) = \frac{2(1-\rho^2)}{(1+\rho^2) - \rho(z+z^{-1})}$$

$$= -2(\rho^{-1} - \rho)\frac{z}{(z-\rho)(z-\rho^{-1})}.$$

The z-plane pole-zero configuration of this function is shown in Figure A.3-5. The overall transformation from $S(\omega)$ to $\mathsf{S}(z)$ is thus given by the replacement

$$\cos\omega \leftarrow \tfrac{1}{2}(z+z^{-1}). \qquad \text{(A.3-6)}$$

For $m \geq 0$ we get

$$R[m] = \text{Res}[\mathsf{S}(z)z^{m-1}; z = \rho]$$

$$= -2(\rho^{-1} - \rho)\frac{\rho\rho^{m-1}}{(\rho - \rho^{-1})}$$

$$= 2\rho^m.$$

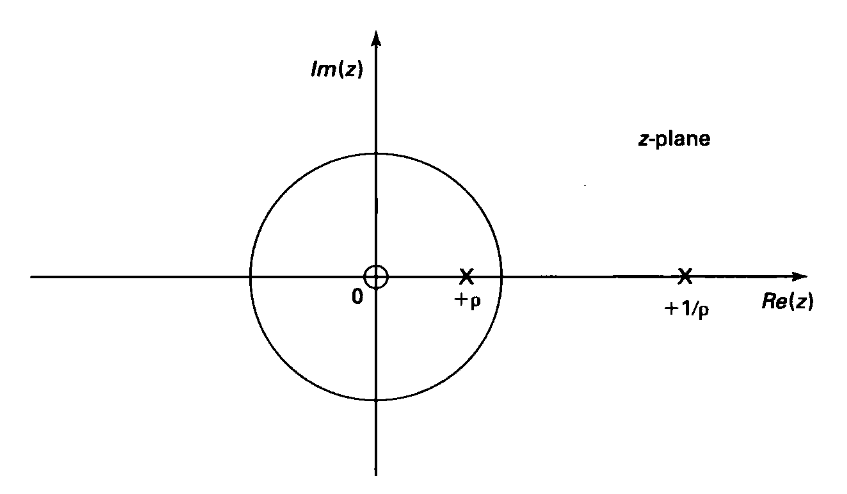


Figure A.3-5 z-plane.

For $m < 0$ we have

$$R[m] = \text{Res}[S(z^{-1})z^{-m-1}; z = \rho],$$

since $z = p^{-1}$ is the one pole outside the unit circle.
Now

$$\begin{aligned} S(1/z) &= -2(\rho^{-1} - \rho)\frac{z^{-1}}{(z^{-1} - \rho)(z^{-1} - \rho^{-1})} \\ &= -2(\rho^{-1} - \rho)\frac{z}{(z - \rho^{-1})(z - \rho)}, \end{aligned}$$

which could easily have been foretold from the symmetry evident in Equation A.3-6. Then

$$\begin{aligned} \text{Res}[S(z^{-1})z^{-m-1}; z = \rho] &= -2(\rho^{-1} - \rho) \left. \frac{z^{-m}(z - \rho)}{(z - \rho^{-1})(z - \rho)} \right|_{z=p} \\ &= -2\frac{(\rho^{-1} - \rho)\rho^{-m}}{(\rho - \rho^{-1})} \\ &= 2\rho^{-m}. \end{aligned}$$

Combining, we get the overall answer

$$R[m] = 2\rho^{|m|}, \qquad -\infty < m < +\infty.$$

A.4 MATHEMATICAL INDUCTION†

Many proofs in probability are obtained by *mathematical induction.* Mathematical induction is a method for obtaining results, especially proving theorems, which are difficult if not impossible to get by any other method. For example: It is claimed that the set S contains all the positive integers. How would we verify this? We could show that $1 \in S$, $2 \in S$, $3 \in S$, etc. But using this procedure would not allow us to finish in finite time. Instead we can use the general *principle of matematical induction*:

Let $\{C_k\}$ be an infinite sequence of propositions, given for all $k \geq 1$. We wish to prove that these propositions are true for every $k \geq 1$. Instead of proving them one by one, we rely on the following principle.

(i) If C_1 is true,
(ii) and for arbitrary $k > 1$, "C_k is true" implies "C_{k+1} is true,"

then C_k holds for all $k \geq 1$.

Thus, we only have to perform the two steps (i and ii), using mathematical induction. After identifying the indexed set of propositions$\{C_k\}$ for our particular problem, we first show that C_1 is true. Then we try to show the second step is true. We do this by assuming that C_k is true for an arbitrary value of positive index k, and then attempting to show that this fact implies that proposition C_{k+1} is true. Then we are finished.

Example A.4-1

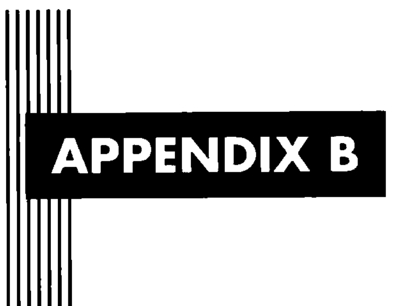
(*mathematical induction*) Show that if $0 < a < b$ then $a^k < b^k$ for all positive integers n.

Solution We choose the method of induction. The problem statement that $0 < a < b$ gives us directly the proposition $C_1 = \{a < b\}$, then we let C_k be the set of positive integers for which $a^k < b^k$, that is, $C_k \triangleq \{a^k < b^k\}$. Now assume that C_k is true, meaning $a^k < b^k$ for some k. It then follows that $a^{k+1} = a \times a^k < a \times b^k < b \times b^k = b^{k+1}$. Thus, C_{k+1} is true. The principle of mathematical induction then allows us to conclude that all the propositions C_k are true, that is, $a^k < b^k$, for all positive integers k.

REFERENCES

A-1. A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* Upper Saddle River, NJ: Prentice Hall, 1989, Chapter 3.
A-2. T. Kailath, *Linear Systems.* Upper Saddle River, NJ: Prentice Hall, 1980, pp. 161–166.
A-3. E. Hille, *Analytic Function Theory,* Vol. I. New York: Blaisdell, 1965, Chapters 7 and 9.
A-4. S. L. Salas and Einar Hille, *Calculus,* 3rd edition, New York: John Wiley & Sons, 1978.

†See [A-4].



Gamma and Delta Functions

B.1 GAMMA FUNCTION

The Gamma function $\Gamma(\alpha)$, for real α, is defined by the integral [1,2]

$$\Gamma(\alpha) \triangleq \int_0^\infty t^{\alpha-1}e^{-t}dt, \tag{B.1-1}$$

where $\alpha > 0$. From Equation B.1-1 we see that $\Gamma(1) = 1$. If we integrate $\Gamma(\alpha+1)$ by parts we obtain

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty t^\alpha e^{-t}dt = -t^\alpha e^{-t}\Big|_0^\infty + \alpha\int_0^\infty t^{\alpha-1}e^{-t}dt \\ &= \alpha\Gamma(\alpha),\end{aligned}$$

Hence, for positive integer k,

$$\Gamma(k) = (k-1)!$$

For values of the argument between the integers, the gamma function does a smooth interpolation. It is available in MATLAB as the function **gamma**.

Therefore, note that $0! = 1$. We leave it to the reader to show that $\Gamma(0.5) = \sqrt{\pi}$ and $\Gamma(1.5) = \sqrt{\pi}/2$. The Gamma function is sometimes called the *generalized factorial function.*

B.2 INCOMPLETE GAMMA FUNCTION

The (upper) incomplete Gamma function $\Gamma(\alpha, x)$ is defined by the integral

$$\Gamma(\alpha, x) \triangleq \int_x^\infty t^{\alpha-1}e^{-t}dt,$$

where $\alpha > 0$. The (lower) incomplete Gamma function is defined by

$$\gamma(\alpha, x) = \int_0^x t^\alpha e^{-t}dt.$$

Unless stated otherwise *incomplete Gamma function* will mean the upper incomplete Gamma function. Clearly $\Gamma(\alpha, 0) = \Gamma(\alpha)$. For $\alpha = k$ an integer, the incomplete Gamma function is known to satisfy the series [3, 4]

$$\Gamma(k, x) = (k-1)!e^{-x}\sum_{l=0}^{k-1}\frac{x^l}{l!},$$

which can also be written as

$$\Gamma(k, x) = (k-1)\Gamma(\alpha-1) + x^{k-1}e^{-x}$$

and it is available in MATLAB as the function `gammainc`. This function plays a crucial role in evaluating the distribution function of the Poisson random variable.

B.3 DIRAC DELTA FUNCTION

The Dirac delta function $\delta(x)$ is often defined as a "function" that is zero everywhere except at $x = 0$, where it is infinite such that

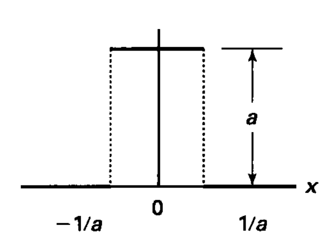
$$\int_{-\infty}^\infty \delta(x)dx = 1.$$

The mild controversy about regarding $\delta(x)$ a "function" in the ordinary sense is partly due to it not being of bounded variation and not having bounded energy in any finite-length support that contains it. Another definition is to regard $\delta(x)$ as the limit of one of several pulses. For example, with rectangular window,

$$w\left(\frac{x}{b}\right) \triangleq \begin{cases} 1, & -b/2 \le x \le b/2, \\ 0, & \text{else,} \end{cases}$$

we can define $\delta(x)$ as

$$\delta(x) \triangleq \lim_{a\to\infty}\{aw(ax)\}.$$



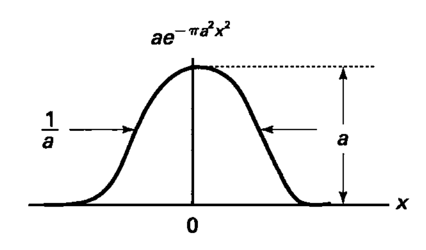


Figure B.3-1 Rectangular and Gaussian-shaped pulses of unit area.

Another possibility is to define $\delta(x)$ as

$$\delta(x) \triangleq \lim_{a\to\infty}\{a\exp(-\pi a^2x^2)\}.$$

The rectangular and Gaussian shaped pulses are shown in Figure B.3-1. The function $aw(ax)$ has discontinuous derivatives, whereas $a\exp(-\pi a^2x^2)$ has continuous derivatives. The exact shape of these functions is immaterial. Their important features are (1) unit area and (2) rapid decrease to zero for $x \neq 0$.

Still another defintion is to call any object a delta function if for any function $f(\cdot)$ continuous at x it satisfies the integral equation†

$$\int_{-\infty}^{\infty} f(y)\delta(y-x)\,dy = f(x). \tag{B.3-1}$$

This definition can, of course, be related to the previous one, since either of the pulses when substituted for $\delta(x)$ in Equation (B.3-1) will essentially furnish the same result when a is large. This follows because the integrand is significantly nonzero only for $x \simeq y$. The integral can, therefore, be approximately evaluated by replacing $f(y)$ by $f(x)$ and moving it outside the integral. Then, since both pulses have unit-area, the result follows. Note that $\delta(x) = \delta(-x)$.

Consider now the unit step $u(x - x_i)$, which is discontinuous at $x = x_i$ with $u(0) \triangleq 1$ (Figure B.3-2a). The discontinuity can be viewed as the limit of the function shown in Figure B.3-2b. The derivative is shown in Figure B.3-2c.

The derivative of the function shown in Figure B.3-2b is given by

$$\begin{aligned}\left.\frac{dF}{dx}\right|_{x_i} &\triangleq \frac{dF(x_i)}{dx_i} = \lim_{\Delta x\to 0}\frac{1}{\Delta x_i}w\left[\frac{x-x_i}{\Delta x}\right] \\ &= \delta(x-x_i).\end{aligned} \tag{B.3-2}$$

†A word of caution is in order here. Since $\delta(x)$ is zero everywhere except at a single point, its integral (in the Riemann sense) is not defined. Hence, Equation B.3-1 is essentially symbolic, that is, it implies a limiting operation as was done with the rectangular and Gaussian pulses.

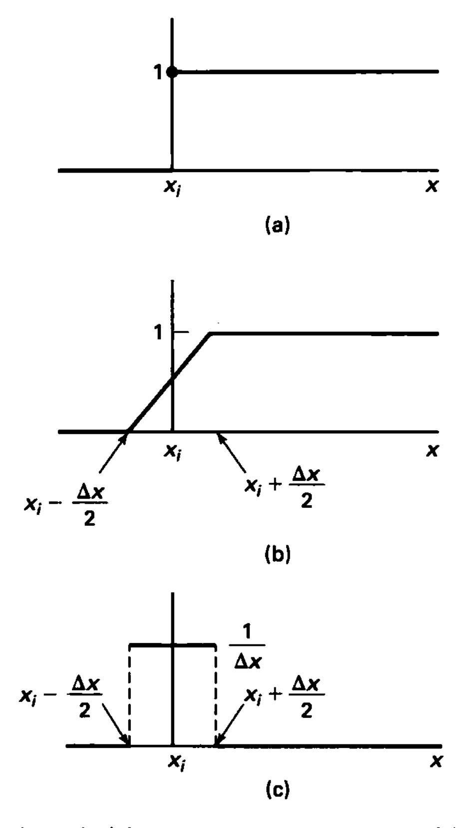


Figure B.3-2 (a) unit step $u(x - x_i)$; (b) approximation to unit step; (c) derivative of function in (b).

Thus, formally, the derivative at a step discontinuity is a delta function with weight† proportional to the height of the jump. It is not uncommon to call $\delta(x - x_i)$ the delta function at "x_i."

Returning now to Equation 2.5-7 in Chapter 2, which can be written as

$$F(x) = \sum_i P_x(x_i)u(x - x_i)$$

and using the result of Equation B.3-2 enables us to write for a discrete RV:

$$f(x) = \frac{dF(x)}{dx} = \sum_i P_X(x_i)\delta(x - x_i), \tag{B.3-3}$$

where we recall that $P_X(x) \triangleq F(x_i) - F(x_i^-)$ and the unit step assures that the summation is over all i such that $x_i \leq x$.

†It is also called the area of the delta function.

REFERENCES

B-1. M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover, 1972. (online at `http://people.math.sfu.ca/~cbm/aands/subj.htm`

B-2. Gamma function Wikipedia page: `http://en.wikipedia.org/wiki/Gamma_function`.

B-3. Incomplete Gamma function page at MathWorld: `http://mathworld.wolfram.com/IncompleteGammaFunction.html`.

B-4. Incomplete Gamma function Wikipedia page: `http://en.wikipedia.org/wiki/Incomplete_gamma_function`.

APPENDIX C Functional Transformations and Jacobians

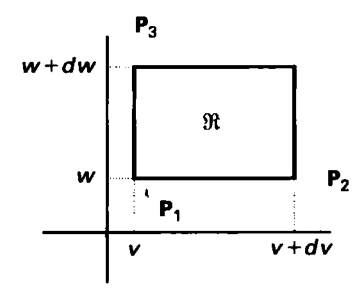
C.1 INTRODUCTION

Functional transformations play an important role in probability theory as well as many other fields. In this appendix, we shall review the theory of Jacobians, beginning with a two-function-to-two-function transformation and extending the result to the n-function-to-n-function case. First, we should recall two basic results from advanced calculus:

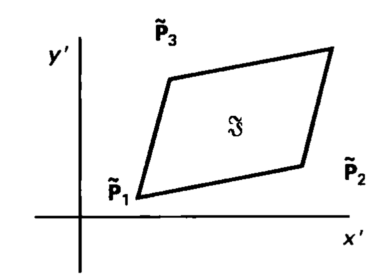
Theorem C.1-1 Consider a bounded linear transformation L from E^n to E^n. If D is a bounded set in E^n with n-dimensional volume $\mathrm{V}(D)$, then the volume of $L(D)$ is merely $\mathrm{k} \times \mathrm{V}(D)$, where k is a constant *independent* of D. ■

Theorem C.1-2 If T is a transformation of class C^1 from E^n to E^n in an open set D then, at every point $p \in D$, dT is a *linear transformation* from E^n to E^n. ■

The first theorem states that the effect of L is merely to multiply the volume by a constant that doesn't depend on the shape of D. The second theorems states that, *at the differential level*, even nonlinear transformations become linear, provided that the transformations consist of differential functions. Both theorems will find application in this development.



Infinitessimal rectangle.



Mapped infinitessimal rectangle into an infinitessimal parallelogram.

Figure C.2-1

C.2 JACOBIANS FOR $n = 2$

Consider the pair of one-to-one[†] differentiable functions $v = g(x,y)$, $w = h(x,y)$ with the unique inverse $x = \phi(v,w)$, $y = \varphi(v,w)$. As the vector $\mathbf{z} = (v,w)$ traces out the infinitesimal rectangle $\Re$ in the v'–w' plane, the vector $\mathbf{u} = (x,y)$ traces out the infinitesimal parallelogram $\Im$ in the x'–y' plane. By Theorem C.1-1, this differential transformation is linear, and by Theorem C.1-2, the ratio of the areas, $A(\Im)/A(\Re)$, is a constant. We shall denote this constant by $|\tilde{J}|$ and compute its value.

We can compute the constant $\tilde{J}$ with the aid of Figure C.2-1. Recalling that $x = \phi(v,w)$, $y = \varphi(v,w)$, we compute the image points $\tilde{\mathbf{P}}_1$, $\tilde{\mathbf{P}}_2$, $\tilde{\mathbf{P}}_3$ of the vertices at $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ as:

$$\tilde{\mathbf{P}}_1 = (x,y),\ \tilde{\mathbf{P}}_2 = \left(x + \frac{\partial\phi}{\partial v}dv, y + \frac{\partial\varphi}{\partial v}dv\right),\ \tilde{\mathbf{P}}_3 = \left(x + \frac{\partial\phi}{\partial w}dw, y + \frac{\partial\varphi}{\partial w}dw\right).$$

These results are directly obtained by a Taylor series expansion about (x,y). Thus, for example, the coordinates (x_2, y_2) of $\tilde{\mathbf{P}}_2$ are obtained from

$$x_2 = \phi(v+dv,w) \approx \phi(v,w) + \frac{\partial\phi}{\partial v}dv \text{ and } y_2 = \varphi(v+dv,w) \approx \varphi(v,w) + \frac{\partial\varphi}{\partial v}dv.$$

There are no nonzero derivatives with respect to w because w is held constant in going from $\mathbf{P}_1$ to $\mathbf{P}_2$. A result from vector analysis, is that the area of a parallelogram spanned by the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is given by the magnitude of the cross-product, that is,

$$A(\Im) = |\mathbf{v}_1 \times \mathbf{v}_2| = \left|\left(\frac{\partial\phi}{\partial v}\mathbf{i}dv + \frac{\partial\varphi}{\partial v}\mathbf{j}dv\right) \times \left(\frac{\partial\phi}{\partial w}\mathbf{i}dw + \frac{\partial\varphi}{\partial w}\mathbf{j}dw\right)\right|,$$

where we used the fact that $\mathbf{v}_1 = \tilde{\mathbf{P}}_2 - \tilde{\mathbf{P}}_1$ and $\mathbf{v}_2 = \tilde{\mathbf{P}}_3 - \tilde{\mathbf{P}}_1$. The unit vectors **i, j** satisfy $\mathbf{i} \times \mathbf{j} = \mathbf{k}, \mathbf{j} \times \mathbf{i} = -\mathbf{k}, \mathbf{i} \times \mathbf{i} = \mathbf{j} \times \mathbf{j} = 0$, where $\mathbf{k} \perp \mathbf{i},\mathbf{j}$ and points out of the plane of the paper. Thus,

[†]This means that every point (x,y) maps into a unique (u,v) and vice versa.

$$A(\Im) = \left| \frac{\partial \phi}{\partial v}\frac{\partial \varphi}{\partial w} - \frac{\partial \phi}{\partial w}\frac{\partial \varphi}{\partial v} \right| dv\, dw.$$

Since $A(\Re) = dv\, dw$, we find that the ratio of the areas is

$$A(\Im)/A(\Re) = \left| \frac{\partial \phi}{\partial v}\frac{\partial \varphi}{\partial w} - \frac{\partial \phi}{\partial w}\frac{\partial \varphi}{\partial v} \right| \triangleq |\tilde{J}|.$$

In higher dimensions it is easier to write $\tilde{J}$ as a determinant. Indeed, even in this two-dimensional case, we can write:

$$\tilde{J} = \begin{vmatrix} \frac{\partial \phi}{\partial v} & \frac{\partial \phi}{\partial w} \\ \frac{\partial \varphi}{\partial v} & \frac{\partial \varphi}{\partial w} \end{vmatrix} = \frac{\partial \phi}{\partial v}\frac{\partial \varphi}{\partial w} - \frac{\partial \phi}{\partial w}\frac{\partial \varphi}{\partial v}.$$

The quantity $\tilde{J}$ is called the Jacobian of the transformation $x = \phi(v, w)$, $y = \varphi(v, w)$.

Among other things, the Jacobian is necessary to preserve probability measure (sometimes called the probability mass or probability volume). For example, consider a pdf $f_{XY}(x, y)$ and the transformation $x = \phi(v, w)$, $y = \varphi(v, w)$. Consider the event $B \triangleq \{\zeta : (X, Y) \in \wp \subset E^2\}$. Then

$$P(B) = \int\int_{\wp} f_{XY}(x, y) dx\, dy \neq \int\int_{\wp} f_{XY}(\phi(v, w), \varphi(v, w)) dv\, dw$$

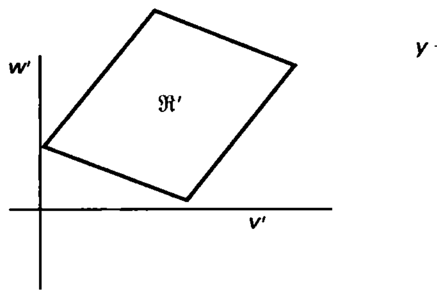
because the volume $dx\, dy \neq dv\, dw$. What is needed is the Jacobian to create the equality among the integrals as

$$\int\int_{\wp} f_{XY}(x, y) dx\, dy = \int\int_{\wp} f_{XY}(\phi(v, w), \varphi(v, w)) |\tilde{J}| dv\, dw.$$

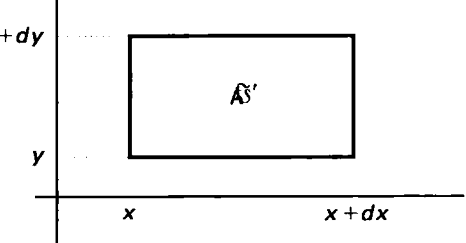
Sometimes it may be easier to deal with the original functions $v = g(x, y), w = h(x, y)$ than the inverse functions $x = \phi(v, w)$, $y = \varphi(v, w)$. To get the desired result, we recompute the ratio of areas by considering the image, $\Re'$, in the v–w system, of an infinitesimal rectangle, $\Im'$, in the x–y system (Figure C.2-2). Following the same procedure as before, we obtain $A(\Im')/A(\Re') \triangleq 1/|J|$, where the primes help indicate the regions in the two systems and J is given by

$$J = \begin{vmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \end{vmatrix}.$$

But, by Theorem C.1-1, $A(\Im')/A(\Re') = A(\Im)/A(\Re)$ and, hence, $|\tilde{J}| = 1/|J|$ or $|\tilde{J}J| = 1$. We leave the details of the computation as an exercise for the reader.



Infinitessimal parallelogram.



Infinitessimal parallelogram is mapped into an infinitessimal rectangle.

Figure C.2-2

C.3 JACOBIAN FOR GENERAL n

The general case is easier to deal with if we allow ourselves to use matrix and vector notation and some results from linear algebra. First, it is not convenient to use the unit vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ in higher dimensions. Instead, we use unit vectors that are represented by column vectors. Thus, in E^2 we use $\mathbf{e}_1 = [1, 0]^T$ and $\mathbf{e}_2 = [0, 1]^T$. Then

$$\mathbf{v}_1 = \frac{\partial \phi}{\partial v} dv\, \mathbf{e}_1 + \frac{\partial \varphi}{\partial v} dv\, \mathbf{e}_2 = \left[\frac{\partial \phi}{\partial v} dv, \frac{\partial \varphi}{\partial v} dv\right]^T$$

and

$$\mathbf{v}_2 = \frac{\partial \phi}{\partial w} dw\, \mathbf{e}_1 + \frac{\partial \varphi}{\partial w} dw\, \mathbf{e}_2 = \left[\frac{\partial \phi}{\partial w} dw, \frac{\partial \varphi}{\partial w} dw\right]^T$$

Next, we form the 2×2 matrix $\mathbf{V}_2 = [\mathbf{v}_1\ \mathbf{v}_2]$, where the subscript 2 on $\mathbf{V}_2$ refers to two-dimensional Euclidean space.

Then, for the special case of $n = 2, A(\Im)$ is given by $|\det \mathbf{V}_2|$. As we go to higher dimensions we drop the term "area of the parallelepiped" in favor of "volume of the parallelepiped," although purists would argue that for spaces of dimensions higher than three we should use "hypervolume." Also in higher dimensions, it is easier to use different subscripts rather than different symbols for functions and arguments. In n-dimensional space, the volume of a parallelepiped is always given by the height times the base area, where the base area is the volume of the parallelepiped in $n-1$ dimensional space and the height is the length of the component of $\mathbf{v}_n$, which is orthogonal to the vectors that span E^{n-1}. Thus, in E^2 the base area is the length of the chosen base vector and the height is the length of the orthogonal component of the second vector. In E^3, the base area is the area of the parallelogram spanned by any two of the three vectors and the height is the length of the component of the third vector orthogonal to the plane containing the first two vectors.

We wish to compute the volume of an infinitesimal parallelepiped in n-dimensional space. Motivated by the fact that the volume, V_2, in two-dimensional space is given by $\mathsf{V}_2 = |\det \mathbf{V}_2|$, we are tempted to write that $\mathsf{V}_n = |\det \mathbf{V}_n|$. Is this true? The answer is yes and the proof is furnished by *induction.* Thus, we assume that $\mathsf{V}_n = |\det \mathbf{V}_n|$ is true and we must prove that $\mathsf{V}_{n+1} = |\det \mathbf{V}_{n+1}|$. Now in terms of the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n, \mathbf{v}_{n+1}$, the matrix $\mathbf{V}_{n+1}$ can be written as

$$\mathbf{V}_{n+1} = \left[\begin{array}{ccc|c} \mathbf{v}_1 & \cdots & \mathbf{v}_n & v_{n+1,1} \\ \vdots & \vdots & \vdots & v_{n+1,2} \\ \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & 0 & v_{n+1,n+1} \end{array}\right]$$

$$= \left[\begin{array}{ccc|c} \mathbf{V}_n & & v_{n+1,1} & \\ & & \vdots & \\ \hline 0 & \cdots & 0 & v_{n+1,n+1} \end{array}\right]$$

To compute $|\det \mathbf{V}_{n+1}|$ we expand by the bottom row to obtain $|\det \mathbf{V}_{n+1}| = |v_{n+1,v+1}||\det \mathbf{V}_n|$, since all other terms in the expansion are zero. Now consider the vector $\mathbf{v}_{n+1}$ in more detail. In terms of the unit vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{n+1}$, it can be written as

$$\mathbf{v}_{n+1} = v_{n+1,n+1}\mathbf{e}_{n+1} + \sum_{i=1}^{n} v_{n+1,i}\,\mathbf{e}_i,$$

where $\mathbf{e}_i$ has a 1 in the ith position (row) and 0's in the remaining n positions. But $\mathbf{e}_{n+1}$ is the unit vector orthogonal to the $\mathbf{e}_1$, $\mathbf{e}_2, \ldots, \mathbf{e}_n$, and hence is orthogonal to the space spanned by them, and $|v_{n+1,n+1}|$ is its height. Also recall that $|\det \mathbf{V}_n|$ is the volume of the parallelepiped in n-dimensions and therefore represents the base area in $n+1$ dimensions. Hence $|\det \mathbf{V}_{n+1}| = |v_{n+1,n+1}||\det \mathbf{V}_n|$ is indeed height times base area and the proof is complete.

Readers familiar with Hadamard's inequality and the Gram–Schmidt orthogonalization procedure can furnish a faster, more direct, proof that avoids induction, but is less intuitive.

Example C.3-1

In Chapter 5 we considered the transformation

$$\begin{aligned} y_1 &= g_1(x_1, x_2, \ldots, x_n) \\ y_2 &= g_2(x_1, x_2, \ldots, x_n) \\ &\vdots \\ y_n &= g_n(x_1, x_2, \ldots, x_n), \end{aligned}$$

with unique inverse

$$x_1 = \phi_1(y_1, y_2, \ldots, y_n)$$
$$x_2 = \phi_2(y_1, y_2, \ldots, y_n)$$
$$\vdots$$
$$x_n = \phi_n(y_1, y_2, \ldots, y_n)$$

Then, a rectangular parallelepiped in the $(y_1, y_2, \ldots, y_n)$ system with volume $\prod_{i=1}^{n} |dy_i|$ maps into a parallelepiped in the $(x_1, x_2, \ldots, x_n)$ system with volume $|\det \mathbf{V}_n| = |\det[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]|$. Here, by computing the differentials of the transformation, we obtain for the $\mathbf{v}_i, i = 1, \ldots, n$:

$$\mathbf{v}_i = \left(\frac{\partial \phi_1}{\partial y_i} dy_i, \ldots, \frac{\partial \phi_n}{\partial y_i} dy_i\right)^T, \qquad i = 1, \ldots, n.$$

APPENDIX D Measure and Probability

D.1 INTRODUCTION AND BASIC IDEAS

Some mathematicians describe probability theory as a special case of *measure theory*. Indeed, random variables are said to be *measurable* functions; the distribution function is said to be a *measure*; events are *measurable* sets; the sample description space together with the field of events is a *measurable space*; and a probability space is a *measure* space. In this appendix, we furnish some results for readers not familiar with the basic ideas of measure theory. We assume that the reader has read Chapter 1 and is familiar with set operations, fields, and sigma fields. The bulk of the material in this appendix is adapted from the classic work by Billingsley.†

Let Ω be a space (a universal set) and let $A, B, C, \ldots$ be elements (subsets) of Ω. Also, as in the text, let ϕ denote the empty set. Let $\Im$ be a field of sets on Ω. Then the pair $(\Omega, \Im)$ is a *measurable space* if $\Im$ is a *σ–field* on Ω. Let μ be a *set function*‡ on $\Im$. Then μ is *a measure* if it satisfies these conditions:

(i) Let $A \in \Im$, then $\mu[A] \in [0, \infty)$;
(ii) $\mu[\phi] = 0$;

†Patrick Billingsley, *Probability and Measure*. New York: John Wiley & Sons, 1978.
‡A set function is a real valued function defined on the field $\Im$ of subsets of the space Ω.

(iii) if $A_1, A_2, \ldots$ is a disjoint sequence of sets in $\Im$ and if $\cup_{k=1}^{\infty} A_k \in \Im$, then

$$\mu\left[\bigcup_{k=1}^{\infty} A_k\right] = \sum_{k=1}^{\infty} \mu[A_k].$$

This property is called *countable additivity.* A measure μ is called *finite* if $\mu[\Omega] < \infty$; it is *infinite* if $\mu[\Omega] = \infty$. It qualifies as a probability measure if $\mu[\Omega] = 1$, as denoted in Chapter 1. If $\Im$ is a σ-field in Ω, the triplet $(\Omega, \Im, \mu)$ is a *measure space.*

Countable additivity implies *finite* additivity, that is,

$$\mu\left[\bigcup_{k=1}^{n} A_k\right] = \sum_{k=1}^{n} \mu[A_k]$$

if the sets are disjoint. A measure μ is *monotone*, that is $\mu[A] \leq \mu[B]$ whenever $A \subset B$. The proof of this statement is straightforward. Write, as is customary in the literature on measure theory, $BA^c \triangleq B - A$ *and* $B = (B - A) \bigcup AB = (B - A) \bigcup A$. Then, since A and B–A are disjoint, it follows that $\mu[B] = \mu[B - A] + \mu[A] \geq \mu[A]$. Also, since $A \bigcup B = (A - B) \bigcup (B - A) \bigcup AB$, it follows that $\mu[A \bigcup B] = \mu[A - B] + \mu[B - A] + \mu[AB]$. This result can be extended to many sets in a σ-field, (sets in a σ-field are called σ-sets), that is,

$$\mu\left[\bigcup_{k=1}^{n} A_k\right] = \sum_{k=1}^{n} \mu[A_k] - \sum_{i<j} \mu[A_i A_j] + \ldots + (-1)^{n+1} \mu[A_1 A \ldots A_n]$$

Of course, this equation makes sense only if the sets have finite measure. It is also straightforward to show that $\mu[\cdot]$ has the property of *subadditivity:*

$$\mu\left[\bigcup_{k=1}^{n} A_k\right] \leq \sum_{k=1}^{n} \mu[A_k].$$

Example D.1-1

Lebesgue measure. Consider the σ-field, $\Im$, of intervals on $\Omega = (0, 1)$. The elements of $\Im$ are called linear *Borel* sets and the σ-field of intervals is called the Borel field $\mathscr{B}$. We shall use this notation for any σ-field on the real line. A measure $\mu[\cdot]$ on $\Im$ is $\mu = \lambda(a, b) \triangleq b - a$, where $b \geq a$. This measure is called the *Lebesgue measure* on $(a, b]$. It can be directly generalized to the real line R^1. An extension of the Lebesgue measure to k-dimensional Euclidean space is:

$$\mu = \lambda_k[x : a_i < x_i \leq b_i, i = 1, \ldots, k] \triangleq \prod_{i=1}^{k} (b_i - a_i)$$

Thus, the Lebesgue measures are *length* $(k = 1)$, *area* $(k = 2)$, *volume* $(k = 3)$, and *hypervolume* $(k > 3)$. We denote the associated σ-field generated by these *generalized rectangles* by the symbol $\mathscr{B}^k$.

There are many important theorems regarding measures. We cite several below.

Theorem D.1-1 (Translation invariance.) Let $A \in \mathscr{B}^k$ and define $A + x \triangleq \{a + x : a \in A\}$. Then $\lambda_k(A + x) = \lambda_k(A)$ for all translation vectors x. ■

Theorem D.1-2 (Lebesgue measure of transformation.) Let $T : R^k \to R^k$ denote a linear and nonsingular transformation from the Euclidean space R^k to R^k. Then $A \in \mathscr{B}^k$ implies that $TA \in \mathscr{B}^k$ and $\lambda_k(TA) = |\det T| \cdot \lambda_k(A)$. For example, if T is a rotation, or reflection, that is, an orthogonal or unitary transformation, then $|\det T| = 1$ and $\lambda_k(TA) = \lambda_k(A)$. ■

Theorem D.1-3 (Lebesgue Measure of Subspaces of R^k). Every $(k-1)$ dimensional hyperplane has k-dimensional Lebesgue measure zero. ■

Theorem D.1-4 (Continuity of measure.) (i) Let μ be a measure on a field $\Im$. Then if A_n and A lie in $\Im$ and $A_n \uparrow A$, then $\mu[A_n] \uparrow \mu[A]$. This is called *continuity of measure from below.* $A_n \uparrow A$ means that $A_{n-1} \subset A_n \subset A_{n+1} \subset \cdots$ *and*

$$A = \bigcup_{n=1}^{\infty} A_n$$

Likewise, $\mu[A_n] \uparrow \mu[A]$ means that $\mu[A_n] \leq \mu[A_{n+1}] \leq \mu[A]$ and $\lim_{n\to\infty} \mu[A_n] = \mu[A]$.

(ii) Let μ be a measure on a field $\Im$. Then if A_n and A lie in $\Im$ and $A_n \downarrow A$, then $\mu[A_n] \downarrow \mu[A]$. This is called *continuity of measure from above.* $A_n \downarrow A$ means that $A_{n-1} \supset A_n \supset A_{n+1} \supset \cdots$ *and*

$$A = \bigcap_{n=1}^{\infty} A_n$$

Likewise, $\mu[A_n] \downarrow \mu[A]$ means that $\mu[A_n] \geq \mu[A_{n+1}] \geq \mu[A]$ and $\lim_{n\to\infty} \mu[A_n] = \mu[A]$. ■

Measurable Mappings and Functions

Let $(\Omega, \Im)$ *and* $(\Omega', \Im')$ be two measurable spaces with two sets $A \in \Im$ and $A' \in \Im'$. For a mapping $T : \Omega \to \Omega'$, consider the inverse image $T^{-1}A' = \{\omega \in \Omega : T\omega \in A'\}$ *for* $A' \subset \Omega'$. The *mapping is measurable* if $T^{-1}A' \in \Im$ *for every* $A' \in \Im'$. For example, consider the unit interval $\Omega = (0, 1)$ with $\Im = \mathscr{B}$ and the mapping $Tx = x^2$. Here, $\Omega' = \Omega$ *and* $\Im' = \mathscr{B}$. Clearly, the inverse image of every Borel interval in Ω' is a Borel interval in Ω. Hence, T is a measurable mapping.

A real function X on Ω, with image space R^1, is said to be *measurable* if its inverse image $X^{-1}B = \{\omega : X(\omega) \in B\} \in \Im$ for every $B \in \Im$.

D.2 APPLICATION OF MEASURE THEORY TO PROBABILITY

A set function P on a σ-field $\Im$ is a *probability measure* if:

(i) $0 \leq P[A] \leq 1$ *for every* $A \in \Im$;
(ii) $P[\phi] = 0$, $P(\Omega) = 1$;

(iii) if $A_1, A_2, \dots, A_k, \dots$ is a disjoint sequence of $\Im$-sets such that

$$\bigcup_{k=1}^{\infty} A_k \in \Im$$

then

$$P\left[\bigcup_{k=1}^{\infty} A_k\right] = \sum_{k=1}^{\infty} P[A_k].$$

(This is the countable additivity property of the probability measure.)

Distribution Measure

In keeping with the notation in the main text, we replace ω with ζ to denote the elements of Ω. Recall that this was done to save ω for the Fourier transform variable needed throughout the text. Let $B \in \mathscr{B}$, the Borel σ-field of intervals on the real line. Consider a (probability) measure μ on $(R^1, \mathscr{B})$ defined by $\mu[B] \triangleq P[\zeta : X(\zeta) \in B] = P_X[B]$. This measure is called the *distribution* or *law* of a random variable. The *distribution function* of X is defined by

$$F_X(x) \triangleq \mu(-\infty, x] = P[X \leq x],$$

where $P[X \leq x]$ is short for $P[\zeta : X(\zeta) \leq x]$. By the *continuity from above* part of the *continuity of measure theorem*, $F_X(x)$ is continuous from the right.

Since the field of events is a *σ-field*, and the distribution function is generated by a measure, all of the properties of measures apply in probability. It is for this reason that probability and measure theories are so closely related. However, to look at probability theory just from the point of view of measure theory is to ignore its rich calculus which enables the solution of engineering, scientific, and statistical problems.



APPENDIX E Sampled Analog Waveforms and Discrete-time Signals

Discrete-time signals are often realized by sampling continuous-time analog wave forms. Here, we briefly review the relationship between the two types of signals. The reconstruction of a continuous-time signal from its equally-spaced samples is governed by the famous Whittaker–Nyquist–Shannon sampling theorem, which states the following.

Theorem E.1-1 *A continuous signal $x(t)$ with real frequencies no higher than $v_{\max}$ can be reconstructed exactly from its samples $x(nT)$ if the sampling interval T satisfies $T < \frac{1}{2v_{\max}}$.* ■

The proof of this important theorem is given in many places, for example, *Principles of Communication Engineering* by John M. Wozencraft and Irwin M. Jacobs, John Wiley and Sons, NY, 1965. Let $x(t)$, $y(t)$, and $h(t)$ denote the input signal, output signal, and impulse response of a linear, shift-invariant (LSI) system respectively. Let B, in Hertz, denote a bandwidth that is greater than any signal or system bandwidth encountered in the system and let $\Delta \triangleq 1/(2B)$. For ease of notation define

$$\text{sinc}\,(x) \triangleq \frac{\sin \pi x}{\pi x}.$$

The relationship between input and output for an LSI system is

$$y(t) = \int_{-\infty}^{\infty} h(s)x(t-s)ds$$

and from the sampling theorem:

$$y(t) = \sum_l y(l\Delta) \text{ sinc } (2B[t - l\Delta]),$$

$$x(t) = \sum_l x(l\Delta) \text{ sinc } (2B[t - l\Delta]),$$

$$h(t) = \sum_l h(l\Delta) \text{ sinc } (2B[t - l\Delta]).$$

If we insert the top three lines into the input–output integral being careful about using different subscripts, and evaluate at $y(t)$ at $t = l\Delta$, we obtain

$$y(l\Delta) = \sum_n \sum_m h(n\Delta)x(m\Delta)I(l, m, n),$$

where

$$I(l, m, n) \triangleq \int_{-\infty}^{\infty} \text{sinc } (2B[s - n\Delta]) \text{ sinc } (2B[s - (l - m)\Delta])ds = 0,$$

for all real integers l, m, n except when $l - m = n$, whereupon it assumes the value Δ. Hence, we obtain the important result that

$$y(l\Delta) = \sum_n h(n\Delta)x\Big([l - n]\Delta\Big)\Delta,$$

Often the factor Δ is submerged into $h(n\Delta)$. In a computer the sampled values of the functions become mere sequences of numbers as $y(l\Delta) \triangleq y[l]$, $x(l\Delta) \triangleq x[l]$, and $h(n\Delta) \triangleq h[n]$. Then, we obtain

$$y[n] = \sum_n h[n]x[l - n]$$

that we recognize as a discrete convolution. The important fact to remember is that the processing of analog signals can be done by operating on their samples and then reconstructing an analog waveform by filtering.

Another point to consider is that the sequence of numbers $\{x[n]\}$ does not contain information about the sampling period, For example, consider the sinusoid $x(t) = A\cos(\omega_r t + \theta)$. If we sample at $t = n\Delta$, $n = \ldots, -2, -1, 0, 1, 2, \ldots$, we obtain the samples $x(n\Delta) = A\cos(n\Delta\omega_r + \theta) = A\cos(n\omega + \theta) \triangleq x[n]$, where $\omega \triangleq \Delta\omega_r$. The radian "frequency" ω is dimensionless, which is consistent with the dimensionless "time" n. It is well to remember that to convert to analog frequencies ω_r (radians/sec) or v_r (Hertz) we must use $\omega_r = \omega\Delta$ or $v_r = v\Delta$. For example, the Fourier transform of a sequence of numbers $\{x[n]\}$ will yield a spectrum of sinusoids at normalized frequencies ω that lie in the interval $[-\pi, \pi]$. If we convert to analog radian frequencies, then the spectrum will lie in the interval $[-2\pi B, 2\pi B]$.

APPENDIX F Independence of Sample Mean and Variance for Normal Random Variables†

Of all the distributions we encounter in probability and statistics, without doubt, the Normal (Gaussian) distribution is of greatest importance. There are a number of reasons for this, but first and foremost is the Central Limit Theorem (CTL), which states that under a set of reasonable and realistic conditions the sum of a large number of independent random variables tends to have a Normal CDF. This property enables us to solve many problems in statistics by invoking the CTL when the sample size is large. Readers of Chapters 6 and 7 will have noticed that we use the CTL to generate results that otherwise would have been difficult to obtain.

There are other reasons why the Normal distribution plays such an important role in probability and statistics. One of them is that the univariate Normal pdf has two parameters that are algebraically independent, that is, within their range they can have any arbitrary values without conflicting with each other. The mean μ can have any value in $(-\infty, \infty)$ and the variance σ^2 can have any value in $(0, \infty)$. This suggests that we can always design a generator of Normal data that will have a specified mean and variance. The same is true for the multivariate, that is, multidimensional, Normal distribution. That is, given a mean vector and covariance matrix, respectively, $\boldsymbol{\mu}, \mathbf{K}$, we can always design a Normal generator whose data will have these parameters. The Normal pdf also enjoys *completeness*, a property of importance in finding a class of optimum estimators called *minimum variance, unbiased estimators.*

Given the importance of the Normal distribution, the estimation of its parameters μ and σ^2 is a central problem in statistics. Assume that we make n i.i.d. observations

†The proof substantially follows that given in [7-1]

on $X:N(\mu,\sigma^2)$. We estimate μ with$(1/n)\sum_{i=1}^{n}X_i$ (sample mean) and and σ^2 with $(1/n)$ $\left(\sum_{i=1}^{n}\left(X_i-(1/n)\sum_{j=1}^{n}X_j\right)^2\right)$ or $(1/(n-1))\left(\sum_{i=1}^{n}\left(X_i-(1/n)\sum_{j=1}^{n}X_j\right)^2\right)$ (sample variance). We note that both the sample mean and sample variance use the same data. Remarkably, the sample mean and sample variance are statistically independent† . In proving this result we shall use a Theorem from probability theory: If the joint moment-generating function of two random variables V and W, say $M_{VW}(t_1,t_2)$, factors as $M_V(t_1)$ $M_W(t_2)$, then V and W are independent. This result was derived in Example 4.7-1 for characteristic functions i.e., moment generating functions evaluated at $t=j\omega$.

The two random variables of interest are the estimators $\hat{\mu}_X$ and $\hat{\sigma}_X^2$. For simplicity and to keep the algebra to a minimum, we define

$$\text{(i) } Y_i \triangleq \frac{X_i-\mu_X}{\sigma_X}, \text{(ii) } V \triangleq n\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^2 = n\hat{\mu}_Y^2$$
$$\text{(iii) } W \triangleq \sum_{i=1}^{n}(Y_i-\hat{\mu}_Y)^2 = (n-1)\hat{\sigma}_Y^2$$

We note in passing that $V:\chi_1^2$ and $W:\chi_{n-1}^2$. Now recall that $M_{VW}(t_1,t_2)$ is given by

$$\begin{aligned} M_{VW}(t_1,t_2) &= E\left[\exp(t_1V+t_2W)\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(2\pi)^{-n/2}\left[\exp\left(-\frac{1}{2}Q\right)\right]dy_1dy_2\cdots dy_n, \end{aligned}$$

where

$$\begin{aligned} Q &\triangleq \sum_{i=1}^{n}y_i^2 - 2\frac{t_1}{n}\left(\sum_{i=1}^{n}y_i\right)^2 - 2t_2\sum_{i=1}^{n}\left(y_i-\left(\frac{1}{n}\sum_{j=1}^{n}y_j\right)\right)^2 \\ &\triangleq \sum_{j=1}^{n}\sum_{i=1}^{n}r_{ij}y_iy_j = \mathbf{y}^T\mathbf{R}^{-1}\mathbf{y} \text{ where } \mathbf{R} \text{ is a covariance matrix} \end{aligned}$$

with diagonal elements r_{ii} and off-diagonal elements r_{ij}, $i\neq j$, where

$$r_{ii} = 1-2t_2-\frac{2(t_1-t_2)}{n}, i=1,...,n, \text{ diagonal terms of } \mathbf{R} \tag{F-1a}$$

$$r_{ij} = -\frac{2(t_1-t_2)}{n}, i,j=1,...,n; i\neq j \text{ off-diagonal terms of } \mathbf{R}. \tag{F-1b}$$

Recalling that the multidimensional Normal pdf is written as

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\mathbf{R}|^{1/2}}\left[\exp(-\frac{1}{2}\mathbf{y}^T\mathbf{R}^{-1}\mathbf{y})\right]$$

† *Independent* or *independence* is meant in a statistical sense. Else we use algebraic or functional independence.

and that $\int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} = 1$, we conclude that

$$\begin{aligned}
M_{VW}(t_1, t_2) &= E\left(\exp(t_1 V + t_2 W)\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(2\pi)^{-n/2}\left[\exp\left(-\frac{1}{2}Q\right)\right] \times dy_1 dy_2 \cdots dy_n. \\
&= |\mathbf{R}|^{-1/2}
\end{aligned}$$

From matrix theory, it is known that for any $n \times n$ matrix $\mathbf{R}$ with diagonal elements a and off-diagonal elements b, the determinant $|\mathbf{R}|$ is computed as $(a-b)^{n-1}(a+(n-1)b)$. Substituting $a \triangleq r_{ii}, b \triangleq r_{ij}$ (from Equation F-1) we obtain

$$\begin{aligned}
M_{VW}(t_1, t_2) &= (1-2t_1)^{-1/2}(1-2t_2)^{-[(n-1)/2]}, t_1 < 1/2, t_2 < 1/2, \\
&= M_V(t_1) \times M_W(t_2).
\end{aligned}$$

Hence by from the Theorem quoted at the beginning of the discussion we conclude that V and W are independent. Hence $F_{VW}(v, w) = F_V(v)F_W(w)$ and therefore that $\hat{\mu}_X^2$ and $\hat{\sigma}_X^2$ are independent. It can be shown that if $\hat{\mu}_X^2$ and σ_X^2 are independent then so are $\hat{\mu}_X$ and σ_X^2. This important result enables us to select separate confidence intervals for $\hat{\mu}_X$ and σ_X^2 without fear of contradiction. The independence of $\hat{\mu}_X$ and σ_X^2 is true only in the Normal case. ■



APPENDIX G Tables of Cumulative Distribution Functions: the Normal, Student t, Chi-square, and F

In the following pages we present tables of the CDF of the (1) Normal; (2) Student-t; (3) Chi-square; and the F, the latter sometimes called the Snedecor F distribution.

The gamma function $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx, \alpha > 0$ appears in several of the CDFs below. When α is an integer, say, $\alpha = m \geq 1$, then $\Gamma(m) = [m-1]! = (m-1) \times (m-2) \times \cdots \times 2 \times 1$. Note 0!=1. Next to each CDF are a few of its applications.

(1) Standard Normal (extensively used in probability and statistics)

$$F_{SN}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{x^2}{2}\right) dx$$

The general univariate Normal CDF is a function of two parameters the mean μ and the variance σ^2.

(2) Student-t (interval estimation, tests on the means of Normal populations $\mu = \mu_0$ versus $\mu \neq \mu_0$)

$$F_T(x;m) = K \int_{-\infty}^{z} \frac{1}{\sqrt{(1+(x^2/m)}} dx, \quad K \triangleq \frac{\Gamma([m+1]/2)}{\Gamma(m/2) \times \sqrt{\pi m}}$$

The Student-t distribution is a function of the parameter m called the *degrees of freedom* (DOF). It is a special case of the F-distribution.

(3) Chi-square (confidence intervals for variance of Normal populations, testing $\sigma^2 = \sigma_0^2$ versus $\sigma^2 \neq \sigma_0^2$, Pearson's goodness-of-fit)

$$F_{\chi^2}(x; m) = K' \int_0^x y^{m/2-1} \exp\left(-\frac{y}{2}\right) dy$$

$$K' \triangleq \frac{1}{2^{m/2}\Gamma(m/2)}$$

The Chi-square CDF is a function of the parameter m called the *degrees of freedom* (DOF).

(4) Snedecor F (generalized likelihood ratio, testing $\sigma_1^2 = \sigma_2^2$ versus $\sigma_1^2 \neq \sigma_2^2$)

$$F_F(x; m, n) = K" \int_0^x y^{m/2-1} \times \left(1 + \frac{my}{n}\right)^{-(m+n)/2} dy$$

$$K" \triangleq \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{m/2}.$$

The Snedecor F CDF is a function of two parameters m and n. These are called the *degrees of freedom* (DOF*) of the F-distribution.* When referring to the DOF, the parameter m is quoted first.

Table 1 Standard Normal CDF

$F_{SN}(x)$ is the table entry. First digit of x gives the row, and second digit of x gives the position in the row.

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table 2 Student-t CDF

For each $F_T(x; n)$ given across the top of the table, row n then determines the table entry, the corresponding value of x.

F / n	0.60	0.75	0.90	0.95	0.975	0.99	0.995	0.9995
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	3.291

Adapted from W.H. Beyer, Ed., in *CRC Handbook of Tables for Probability and Statistics*, 2d ed., The Chemical Rubber Co., Cleveland, 1968; p. 283. With permission of CRC Press, Inc.

Table 3 Chi-Square CDF

For each $F_{\chi^2}(x;n)$ given across the top of the table, row n then determines the table entry, the corresponding value of x.

n/F	.005	.010	.025	.050	.100	.250	.500	.750	.900	.950	.975	.990	.995
1	$.0^4393$	$.0^3157$	$.0^3982$	$.0^2393$	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7

This table is abridged from "Tables of Percentage Points of the Incomplete Beta Function and of the Chi-square Distribution," *Biometrika* Vol. 32 (1941). It is here published with the kind permission of the author, Catherine M. Thompson, and the editor of *Biometrika*.

Table 4 Cumulative F CDF

For each n in the second column on the left and each m in the uppermost row, the entry in the table furnishes the argument needed to yield $F_F(x; m, n)$ in the column at the extreme left.

G	n	m 1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
.90		39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.7	61.2	61.7	62.3	62.8	63.1	63.3
.95		161	200	216	225	230	234	237	239	241	242	244	246	248	250	252	253	254
.975	1	648	800	864	900	922	937	948	957	963	969	977	985	993	1000	1010	1010	1020
.99		4,050	5,000	5,400	5,620	5,760	5,860	5,930	5,980	6,020	6,060	6,110	6,160	6,210	6,260	6,310	6,340	6,370
.995		16,200	20,000	21,600	22,500	23,100	23,400	23,700	23,900	24,100	24,200	24,400	24,600	24,800	25,000	25,200	25,400	25,500
.90		8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.46	9.47	9.48	9.49
.95		18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
.975	2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5
.99		98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5
.995		199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199
.90		5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.15	5.14	5.13
.95		10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.55	8.53
.975	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.0	13.9	13.9
.99		34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.5	26.3	26.2	26.1
.995		55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9	43.7	43.4	43.1	42.8	42.5	42.1	42.0	41.8
.90		4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.93	3.92	3.90	3.87	3.84	3.82	3.79	3.78	3.76
.95		7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.66	5.63
.975	4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.46	8.36	8.31	8.26
.99		21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.7	13.6	13.5
.995		31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1	21.0	20.7	20.4	20.2	19.9	19.6	19.5	19.3
.90		4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.17	3.14	3.12	3.11
.95		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.40	4.37
.975	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.23	6.12	6.07	6.02
.99		16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.38	9.20	9.11	9.02
.995		22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8	13.6	13.4	13.1	12.9	12.7	12.4	12.3	12.1
.90		3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.80	2.76	2.74	2.72
.95		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.70	3.67
.975	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.07	4.96	4.90	4.85
.99		13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.97	6.88
.995		18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4	10.2	10.0	9.81	9.59	9.36	9.12	9.00	8.88

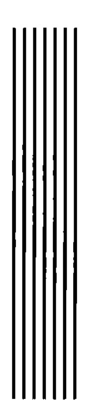
G	n	m	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
.90			3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.56	2.51	2.49	2.47
.95			5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.27	3.23
.975	7		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.36	4.25	4.20	4.14
.99			12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.74	5.65
.995			16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.53	7.31	7.19	7.08
.90			3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.38	2.34	2.31	2.29
.95			5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01	2.97	2.93
.975	8		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.89	3.78	3.73	3.67
.99			11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.95	4.86
.995			14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.40	6.18	6.06	5.95
.90			3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.25	2.21	2.18	2.16
.95			5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.75	2.71
.975	9		7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.56	3.45	3.39	3.33
.99			10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.40	4.31
.995			13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.62	5.41	5.30	5.19
.90			3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.15	2.11	2.08	2.06
.95			4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.70	2.62	2.58	2.54
.975	10		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.31	3.20	3.14	3.08
.99			10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	4.00	3.91
.995			12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	5.07	4.86	4.75	4.64
.90			3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.01	1.96	1.93	1.90
.95			4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.47	2.38	2.34	2.30
.975	12		6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.96	2.85	2.79	2.72
.99			9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.70	3.54	3.45	3.36
.995			11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.33	4.12	4.01	3.90
.90			3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.87	1.82	1.79	1.76
.95			4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.11	2.07
.975	15		6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.64	2.52	2.46	2.40
.99			8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.96	2.87
.995			10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.69	3.48	3.37	3.26

(Continued)

Table 4 (Continued)

G	n	m 1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	120	∞
.90		2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.74	1.68	1.64	1.61
.95		4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.90	1.84
.975	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.35	2.22	2.16	2.09
.99		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.52	2.42
.995		9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	3.12	2.92	2.81	2.69
.90		2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.61	1.54	1.50	1.46
.95		4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.68	1.62
.975	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.07	1.94	1.87	1.79
.99		7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.11	2.01
.995		9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.63	2.42	2.30	2.18
.90		2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.48	1.40	1.35	1.29
.95		4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.47	1.39
.975	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.82	1.67	1.58	1.48
.99		7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.73	1.60
.995		8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.74	2.57	2.39	2.19	1.96	1.83	1.69
.90		2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.54	1.48	1.41	1.32	1.26	1.19
.95		3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.43	1.35	1.25
.975	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.69	1.53	1.43	1.31
.99		6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.66	1.53	1.38
.995		8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81	2.71	2.54	2.37	2.19	1.98	1.75	1.61	1.43
.90		2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.34	1.24	1.17	1.00
.95		3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.32	1.22	1.00
.975	∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.57	1.39	1.27	1.00
.99		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.47	1.32	1.00

*This is table is abridged from "Table of percentage points of the inverted beta distribution," *Biometrika*, Vol. 33 (1943). It is here published with the king permission of authors, Maxine Merrington and Catherine M. Thompson, and the editor of *Biometrika*.



Index